Examining Power and Type 1 Error for Step and Item Level Tests of Invariance: Investigating the Effect of the Number of Item Score Levels

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Alicia Nicole Ayodele

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ernest C. Davenport, Jr.

May 2017

Acknowledgements

First, I would like to thank my adviser, Ernest C. Davenport, Jr. for his feedback, encouragement, and patience with me as I persevered to finish this degree. This has been quite a long journey for both of us. Thank you for believing in me and helping me finish.

Second, I would like to thank my committee members, Mark L. Davison, Michael R. Harwell, and Adam J. Rothman. Your honest feedback and breadth of knowledge made me a better researcher. Thank you for taking the time to help me complete this process.

Third, I would like to thank Doneka Scott, Michelle Kuhl, Frances Lawrenz, and Theodore Christ who have all given me not only advice, but also funding to help complete my degree. You have all been a key part of my journey. Additionally, it was very helpful to leave the program without having the financial burden of student loans.

Fourth, I would like to thank my family. My parents, Arnold and Adwina Baptiste, instilled in me the importance of hard work and education and encouraged me to achieve what I did not think I could. Life can be difficult, but we are still standing--I love you both. My husband, Michael Ayodele, you are the love of my life, my teammate, partner, biggest cheerleader, and best friend. You encouraged me at my lowest points, sacrificed for me, and believed in me when I wanted to give up. You mean the world to me and I am blessed to have you.

Finally, I have to thank God. So much has happened since I first walked into the QME department, including becoming a mother of three beautiful children. Without my faith, I could not stand. I am so grateful.

Dedication

This thesis is dedicated to my children: Daniel, Brandon, and Neya Ayodele.

Abstract

Within polytomous items, differential item functioning (DIF) can take on various forms due to the number of response categories. The lack of invariance at this level is referred to as differential step functioning (DSF). The most common DSF methods in the literature are the adjacent category log odds ratio (AC-LOR) estimator and cumulative category log odds ratio estimator (CU-LOR). Although the study of DSF may be helpful when opposing DIF effects within an item can go undetected or for informing what part of a multi-step item may need improvement, research regarding DSF procedures is limited. The effect of number of item score levels has not been investigated with regard to the relationship between DSF and traditional DIF methods, including differences in statistical behavior. This study investigates the effect of the number of item score levels on power and Type I error of the following DSF methods: AC-LOR, CU-LOR as well as DIF methods: Mantel (chi-square) Test, Liu Agresti, Generalized Mantel-Haenszel, and Simultaneous Step Level test (SSL). This study also examined which statistical procedures are most effective for adjusting per comparison Type I errors for the SSL method: Dunn-Bonferroni, Benjamini and Hochberg, or Holm's. Conditions varied included (a) sample size ratio, (b) number of item score levels, (c) generating model, (d) impact, and (e) DSF pattern. Results suggest that altering the number of score levels did not have an effect on the DSF/DIF detection methods. When considering both statistical and practical significance of factors affecting power, the pattern of DSF was the most important effect. Additionally, the Dunn-Bonferroni adjustment was adequate when using the SSL method. The SSL method performed well compared to the other DIF methods and should be considered for simultaneously detecting both DSF and DIF. The significance of these results as well as limitations and future directions are discussed.

TABLES OF CONTENTS

LIST OF TABLES.....vii

LIST OF FIGURES..... x

CHAPTER 1..... 1

INTRODUCTION..... 1

CHAPTER 2..... 5

LITERATURE REVIEW..... 5

Differential Item Functioning Terminology..... 5

Parametric versus Nonparametric Polytomous DIF and DSF Tests..... 12

Description of Differential Step Functioning Detection Procedures..... 14

Adjacent Category Approach..... 15

Cumulative Category Approach..... 15

Comparison to Other DSF Methods..... 17

Description of Polytomous DIF Detection Procedures..... 18

Mantel Test..... 18

Liu-Agresti..... 21

Simultaneous Step Level Test....22

Comparison to Other Observed Score Approach DIF Methods....23

Addressing Inflated Type I errors in DIF/DSF Test for Polytomous Items....25

Dunn-Bonferroni....26

Benjamini and Hochberg....26

Holm....27

Other Methods for Addressing Inflated Type I Error Rates....28

Factors Considered in DIF/DSF Simulation Studies....29

Sample Characteristics....29

Test Characteristics....30

Analysis Characteristics....31

DIF/DSF Detection in Applied Research....34

Summary of Literature Review....36

CHAPTER 3....38

METHODS....38

Data Generation....39

Study Conditions....40

Analysis....46

CHAPTER 4....49

RESULTS....49

Research Question 1: Type I Error Rates and Tests of Invariance....49

Research Question 2: Type I Error Adjustments for Multiple Significance Testing....66

Research Question 3: Effect of Conditions on Power (ANOVA)....70

CHAPTER 5....85

DISCUSSION....85

Synthesis of Findings....85

Limitations and Future Research....91

REFERENCES....93

APPENDIX A: EXAMPLE MULTIPLE SCORE LEVEL ITEM....106

APPENDIX B: P-VALUE ADJUSTMENT GRAPHS....107

APPENDIX C: R FUNCTIONS FOR DATA GENERATION AND CREATION OF DIF/DSF DETECTION METHODS....109

LIST OF TABLES

Table 1 . Comparison of Classifications for example DSF/DIF Item Effects..... 11

Table 2 . The *k*th level of a 2 x J contingency table.....15

Table 3 . Parameters for items with three score levels, convergent DSF condition..... 43

Table 4 . Parameters for items with four score levels, convergent DSF condition.....44

Table 5 . Parameters for items with three score levels, divergent DSF condition.....45

Table 6 . Parameters for items with four score levels, divergent DSF condition..... 46

Table 7 . Step level rejection rates of studied items, no impact, no DSF..... 52

Table 8 . Step level rejection rates of studied items with no impact and convergent DSF53

Table 9 . Step level rejection rates of studied items with no impact and divergent DSF.. 55

Table 10 . Type 1 Error Rates of DSF and DIF Detection Methods for Items with No DSF56

Table 11 . Type 1 Error Rates of DSF and DIF Detection Methods for Items with Convergent DSF.....57

Table 12 . Type 1 Error Rates of DSF and DIF Detection Methods for Items with Divergent DSF..... 59

Table 13 . Power Rates of DSF and DIF Detection Methods for Items with Convergent DSF..... 60

Table 14 . Power Rates of DIF Detection Methods for Items with Divergent DSF..... 61

Table 15 . Item level rejection rates of studied items with no impact and no DSF..... 63

Table 16 . Item level rejection rates of studied items with no impact and convergent DSF64

Table 17 . Item level rejection rates of studied items with no impact and divergent DSF 65

Table 18 . Adjusted Type 1 Error and Power Rates for the Simultaneous Step Level (SSL) DIF test with No DSF*..... 67

Table 19 . Adjusted Type 1 Error and Power Rates for the Simultaneous Step Level (SSL) DIF test with Convergent DSF*.....68

Table 20 . Type 1 Error and Power Rate Adjustments for the Simultaneous Step Level (SSL) DIF Test with Divergent DSF..... 69

Table 21 . ANOVA results for statistical power rates using adjacent category log odds ratio (AC-LOR).....71

Table 22 . ANOVA results for statistical power rates using cumulative category log odds ratio (CU-LOR).....72

Table 23 . ANOVA results for statistical power rates using the Mantel Test.....73

Table 24 . ANOVA results for statistical power rates using the Generalized Mantel Haenszel statistic (GMH).....74

Table 25 . ANOVA results for statistical power rates using the Liu-Agresti statistic.....75

Table 26 . ANOVA results for statistical power rates using the Simultaneous Step Level (SSL) Test..... 76

Table 27 . Means and Standard Deviations for Statistical Power using Adjacent Category Log Odds Ratio (AC-LOR).....77

Table 28 . Means and Standard Deviations for Statistical Power using Adjacent Category Log Odds Ratio (AC-LOR).....79

Table 29 . Means and Standard Deviations for Statistical Power using Cumulative Log Odds Ratio (CU-LOR)..... 80

Table 30 . Means and Standard Deviations for Statistical Power using the Generalized Mantel Haenszel Test (GMH).....81

Table 31 . Means and Standard Deviations for Statistical Power using the Mantel Test.. 82

Table 32 . Means and Standard Deviations for Statistical Power using the Liu-Agresti Statistic.....83

Table 33 . Means and Standard Deviations for Statistical Power using the Simultaneous Step Level Test.....84

LIST OF FIGURES

Figure 1 . Patterns of Differential Item Functioning for Two Score Levels/Categories..... 8

Figure 2 . Cumulative category log odds ratio (CU-LOR) *p*-values for selected conditions..... 107

CHAPTER 1

INTRODUCTION

To identify the level of a trait across groups or examine differential correlates, one must assume numerical values for groups are on the same measurement scale, "measurement invariance" (Drasgow, 1984, 1987); otherwise, differences may be misleading (Reise & Widaman, 1993). The validity of a test must come into question if it is not accurately measuring the construct equally across all test takers; particularly high stakes tests which influence educational placement of students. Recent studies have explored within item invariance for polytomous items; the lack of invariance at this level is referred to as differential step functioning (DSF). (Gattormorta, Penfield, & Myers, 2012; Penfield, 2007, 2008; Miller, Chahine, & Childs, 2010). With polytomous items, differential item functioning (DIF) can take on various forms due to the number of response categories. For J step functions and r response categories, the conditional probabilities associated with each response category are derived through the parameterization of $J = r - 1$ step functions (Penfield, 2007). A particular step function is denoted j, such that $j = 1, 2, \ldots J$. Due to the between-group differences in the measurement properties that can vary in magnitude and/or sign across the J steps, DIF can take on various forms at each of the J steps. A between group difference in the measurement properties at a particular step in a polytomous item is what characterizes DSF (Penfield, 2007). It may be helpful to investigate DSF if differential item functioning (DIF) effects that are opposite in sign or magnitude cause item bias to go undetected (Penfield 2007). Additionally, DSF effects give more information on which parts of a multistep item may need improvement. The most common DSF methods in the literature are the adjacent category log odds ratio (AC-LOR) estimator and cumulative category log odds ratio estimator (CU-LOR).

Examples of DSF studies include Gattormorta, Penfield, and Myers (2012) who found DSF effects in a School Indicators Achievement Program assessment that consisted

of 30 polytomous math items. Students who took the French version of the test were favored at one score level while students who took the English version were favored at another. In another example, Miller, Chahine, and Childs (2010) combined DIF and DSF methods to investigate the effect of teacher instructional practices on a 9th grade assessment of mathematics. Miller et al. (2010) interpreted DSF effects at the lower score levels as indicators of conceptual understanding which is more likely to be impacted by instruction. The results were inconclusive due to unreliability of teachers self-reporting their instructional practices and emphasized the need for more research on these methods to not only inform item development, but also instructional practices. Penfield (2007) investigated the power and Type I error of DSF estimators under various conditions (i.e. differences in generating model, ability distribution mean, type of DSF), finding that DSF estimators were more powerful and accurate than the omnibus DIF approach, but recommended that future research investigate the number of response options, different patterns of DSF for the studied item(s), less conservative Type I error adjustments, and group size effect.

Lack of invariance at the score level influences DIF results at the item level; additionally, the pattern of DIF at the score level may have more of an effect than the percent of items that lack invariance for some DIF estimators, such as Mantel-Haenszel procedures (Wang & Su, 2004). Atar (2007) investigated three different DIF procedures (Likelihood Ratio test, generalized linear model based test, and logistic regression) for mixed (dichotomous and polytomous) tests and found that, although the combination of sample size and DIF magnitude had the most effect on power and type I error of DIF detection, for some procedures, this effect was moderated by which score level threshold exhibited DIF. Atar (2007) noted that large scale assessments are likely to have polytomous items with differing number of score categories (i.e. the Florida Comprehensive Achievement Test –FCAT – has items with three short response categories and items with five extended response categories) and that this effect should be considered in future research.

Multiple significance tests are required for detecting lack of invariance for

polytomous items. Penfield (2007) suggested investigating a less conservative approach than the Bonferroni correction for adjusting the per comparison Type I error rate for DSF estimators and the Simultaneous Step Level test of DIF. Kim (2010) investigated the Type I error rate of several parametric (Item Response Theory) procedures as well as Mantel-Haenszel and logistic procedures. It was discovered that the Benjamini and Hochberg procedure, when compared to Bonferroni and Holm's procedures, performed best at controlling the Type I error rate while maintaining adequate power for detection of DIF in dichotomous items where multiple significance tests were used, but suggested that findings should be generalized to polytomous items in future studies.

Based on findings in the literature, this study will investigate the effects of polytomous item features on the power and Type I error of nonparametric invariance tests of differential step and item functioning. Nonparametric tests are of interest due to the restrictions that parametric tests may impose, particularly sample size and model fit (Penfield, 2008). The effect of number of item score levels has not been investigated with regard to the relationship between DIF and DSF and therefore will be the main condition investigated among the following methods: DSF tests of invariance--adjacent category and cumulative log odds ratio estimators--and DIF tests of invariance--Mantel (chi-square) Test, Liu Agresti, Generalized Mantel-Haenszel [GMH], and Simultaneous Step Level test [SSL]. These nonparametric methods were chosen because a) research on DSF is minimal thus far (Penfield, 2007); b) there are fewer studies on the performance of Mantel-Haenszel based DIF procedures for polytomous items than for dichotomous items (Wang & Su, 2004); c) Penfield (2009) suggested at the time that the relative performance of SSL test to other procedures has not been fully explored; and d) based on Kim (2010) and Penfield's (2007) recommendations for investigating Type I error adjustments in polytomous DIF, this study will examine which statistical procedures are most effective for adjusting per comparison Type I errors for DIF detection in polytomous items: Bonferroni, Benjamini and Hochberg, or Holm's.

The research questions are:

- In terms of Type I error and power, which non-parametric test of invariance performs

better as the number of response categories increase for polytomous items among the following DSF tests of invariance (adjacent category and cumulative log odds ratio estimators) or DIF tests of invariance (Mantel Test, Liu Agresti, Generalized Mantel-Haenszel, Simultaneous Step Level test)?

- Of the Dunn-Bonferroni, Benjamini and Hochberg, and Holm's methods, which procedure works best for controlling Type I error in the SSL method due to multiple significance tests of DIF for polytomous items?
- Do differences in the generating model, ability distributions, and pattern of DSF affect the power of DSF and DIF tests as the number of score levels increase?

CHAPTER 2

LITERATURE REVIEW

The purpose of this literature review is to provide a summary and critical review of terminology/methodology for differential item functioning (DIF) and differential step functioning (DSF). First, the review will discuss key definitions used in the DIF and DSF literature. The next section describes methods used to detect differential item and step functioning. Following this will be a discussion of ways to address Type I error for the methods described. Finally, the literature review will offer a comparison of findings and limitations in simulation as well as applied studies..

Differential Item Functioning Terminology

Test scores are *invariant* across groups when items on a test measure the same trait in a way that provides comparable scores for the two or more groups such that numerical designations for groups are on the same measurement scale (Reise, Widaman, & Pugh 1993; Thissen, Steinberg, & Girrard, 1986). Another way to think about it is that persons with the same level of ability on a trait, but from different groups, have the same probability of success on an item (Swaminathan & Rogers, 1990). Thus, when a test is invariant across populations, mean differences are only reflective of true mean differences between the populations on the trait (Raju, Laffitte, & Byrne, 2002). Otherwise, comparisons across groups are inaccurate and meaningless which could be detrimental when making high stakes decisions.

While the literature primarily uses the term measurement invariance to discuss confirmatory factor scale invariance across multiple groups, item response theory (IRT) based methods give more information at the item level, particularly for categorical items, and uses terms such as *item bias* and *differential item functioning* (Camilli & Shepard, 1994; Chang, Mazzeo & Roussos,1993; Dorans & Schmidt 1991; Swaminathan & Rogers, 1990; Zwick, 1990; Smith, 2004; Woods, 2008). Concerning items that make up the test, bias is generally due to “construct underrperesentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers” (Standards for Educational and Psychological Testing, 2014). Differential item functioning refers to a statistical property “in which different groups of test takers who

have the same total test score have different average item scores, in some cases, different rates of choosing various item options" (Standards for Educational and Psychological Testing, 2014).

Differential step functioning is a term first used by Penfield (2007) to describe "a between group difference in measurement properties within a particular step of a polytomous item". Differential step functioning can be viewed as a subset of differential item functioning that focuses on within item DIF effects, called DSF effects, as opposed to a single DIF effect for a polytomous item. At the time this literature review was conducted, less than ten articles were found that addressed this issue. Key terms in this section used to describe a DIF analysis are synonymous with terms used for a DSF analysis, with the exception that the focus of DSF are, again, the within item effects for a polytomous item.

Typically a *reference* and *focal* group are identified for conducting a DIF analysis; it is understood that the reference group is the favored or majority group based on previous knowledge from studies with similar sample demographics (Holland & Thayer 1988; Shealy & Stout, 1993; Woods, 2008). Comparison groups observed most frequently in the literature include ethnicity, gender, native language spoken, test accommodation group, or school attending (e.g. Gattamorta, Penfield, and Myers, 2012; Gilmore, 2014; Reise et al., 1993; Steinberg & Thissen, 2006; Swaminathan & Rogers, 1990; Taylor & Lee, 2012; Thissen et al., 1986). When the construct is inadequately or inaccurately defined across groups, it is possible to have either identifiable or unintentional "other traits" (multidimensionality) in at least one of the groups (Camilli, 2006; Hattie, 1985; Thissen et al., 1986). Multidimensionality may cause different small, unique effects on items in one or more groups even if the test appears to be unidimensional within each group which may also lead to bias (Thissen et al., 1986).

With respect to the observed true score, DIF can be identified in an item--referred to as the *studied* item--if there is one value of the observed true score for which its expectation is different for one group versus the other (Chang, Mazzeo, & Roussos, 1996; Shealy & Stout 1993). Items assumed to be free of DIF are referred to as *anchor* items (Chang, Mazzeo, & Roussos, 1996; Shealy & Stout 1993). To conduct the analysis, examinees can be matched on an internal or external criteria, the *matching variable*

(Donaghue & Allen, 1993; Zwick, Donaghue & Grima, 1993). Typically the matching variable is the total test score composed of anchor items, although Zwick et al. (1993) proposed that including the studied item promotes stability in DIF detection for Mantel-Haenszel based procedures. For some cases, it may be necessary to preserve the quality of the matching variable if too many items are identified as exhibiting DIF. When too many items exhibit DIF, the matching variable becomes contaminated and may be inappropriate to use for comparing groups when conducting a DIF analysis. Thus, *purification* may be implemented such that DIF items are systematically removed from the matching variable, while analyzing all items for DIF (Park & Lautenschlager, 1990; Wang & Su, 2004).

In IRT, item response functions describe the relationship between item responses and the underlying latent variable which is the educational/psychological construct being measured (McDonald, 1999; Steinberg, 2001). DIF is present when item response functions (IRF) for an item differ for each group when examinees are matched by latent ability (Chang, et al., 1996; Steinberg & Thissen, 2006; Steinberg, 2001; Thissen et al., 1986). In the IRT framework, DIF can be expressed in the threshold parameter (denoted as *b*) of the item response function and represents the endorsement rate for each group; a significant shift in magnitude of *b* between groups is the simplest way that DIF can occur (Steinberg, 2001; Wood, 2011). *Uniform* DIF refers to bias in favor of one group that is constant across ability levels; *non-uniform* DIF refers to bias in favor of one group that is not constant across ability levels (Hambelton & Rogers, 1989; Mellenberg, 1982; Penfield, 2010; Swaminathan & Rogers, 1990). A few authors have suggested the term *unidirectional* bias as a general description of both uniform and non-uniform DIF (Hanson, 1998; Shealy & Stout, 1993). Some IRT models include the value for which items discriminate between examinees with different ability/trait levels (denoted as the slope of the IRF or the *a* parameter e.g. GRM; Samejima 1969, 1972). DIF within this parameter is referred to as *crossing* DIF in dichotomous items and suggests that item discrimination is different across groups (Camilli & Shepard, 1994; Finch & French, 2007; Li & Stout, 1996; Penfield & Algina, 2003; Penfield & Camilli, 2007; Swaminathan & Rogers, 1990; Thissen, Steinberg, & Wainer, 1993). For DSF, the response function that defines the transition from category J to category $J + 1$ is also

called a step function, and can be parameterized differently depending on the IRT model (i.e. inclusion of additional parameters besides the *b* parameter). Crossing DSF effects within polytomous items has only been studied within the last five years; it can occur either within the *a* parameter or as the result of different combinations of DSF effects within item step functions (Penfield 2010a, 2010b). Figure 1 shows an example of uniform, non-uniform, and crossing DIF patterns.

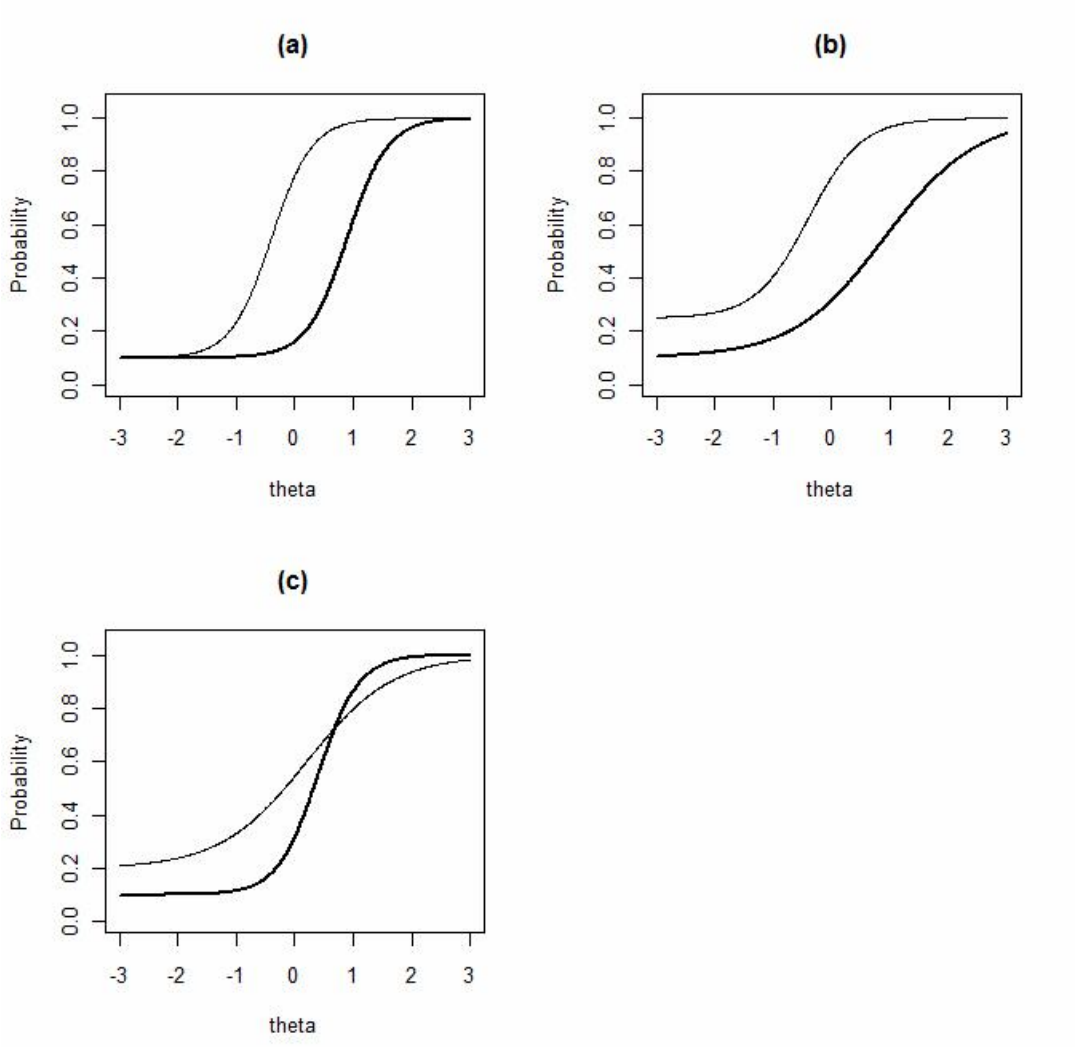


Figure 1. Patterns of Differential Item Functioning for Two Score Levels/Categories. For the range of ability (theta), the reference group item characteristic curve is the thick line; the focal group item characteristic curve is the thin line. Patterns include (a) uniform DIF, (b) non-uniform DIF; and (c) crossing DIF.

While the descriptions of DIF mentioned (uniform, non-uniform, crossing) are adequate to account for DIF in dichotomous items, additional patterns are exhibited in polytomous items due to the multiple category response functions (CRF) or step functions within the item. The labeling of these patterns have varied within the literature. DIF patterns within a polytomous item may be labeled *high-shift* (occurs at the last CRF); *low shift* (occurs at the first CRF); *single level* (occurs within any single CRF); *variable* (occurs within more than one, but not all CRFs); d*ivergent/divergent-1/balanced* (DIF favors both groups and cancels out); *divergent-2* (DIF favors both groups but does not cancel); *constant* (same magnitude of DIF across CRFs); and finally *none*/*no* DIF (Ankenmann, Witt, & Dunbar, 1999; Atar, 2007; Chang, et al. 1996; Penfield & Algina 2003; Penfield 2007, Penfield, 2010b; Wang & Su 2004; Zwick, 1993). With respect to magnitude, studies commonly use the ETS effect size classification for the MH statistic. The MH statistic is derived from odds ratios. The log odds is typically used to determine effect size due to its symmetry about zero such that positive values indicate DIF in favor of the reference group and negative values indicate DIF in favor of the focal group. An effect less than the absolute value of 0.43 is a small effect, an effect between 0.43 and 0.64 in absolute value is medium, and an effect greater than 0.64 in absolute value is a large effect. (Penfield, 2007). As an example, for two groups of the same ability, let us say that the difficulty of a multi-step math problem with no DIF at step one is measured at 2 on a logit scale (natural log odds of getting it correct). A small DIF effect could mean that this item would be +.3 harder or 2.3 for the focal group versus 2 for the reference group. Consequently, a medium DIF effect could mean the item is measured at 2.6 for the focal group and a large effect could mean that the item is measured at 2.7 for the focal group versus 2 for the reference group. Thus, even though the groups have the same level of ability, the word problem is harder for the focal group than for the reference group. This could be the case, for instance, if the focal group members are not native English speakers who have difficulty when met with word problems as opposed to numeric equations. An example math problem with multiple score levels is given in Appendix A.

A two dimensional taxonomy for DSF has been proposed by Penfield (2008). For

the DSF taxonomy, the first dimension concerns the pervasiveness of DSF. For example, if moderate to large magnitudes of DSF effects are present at all item steps, DSF is *pervasive*; otherwise it is *non-pervasive*. The second dimension of the DSF taxonomy describes the pattern of DSF: *constant* (DSF favors one group at the same magnitude), *convergent* (DSF favors one group at varying magnitudes), *divergent* (where the reference group is favored in one step and the focal group in the other, for instance). Although Wood (2011) referred to Penfield's (2008) DSF taxonomy as a DIF taxonomy, these should be viewed as separate. For instance, Penfield (2008) would refer to a 4 category item exhibiting small DSF (i.e. 0.3) at every item step as having no DSF because small magnitudes of DSF do not count towards the pattern. However, this same item could be viewed as exhibiting large DIF (0.9) with a constant uniform pattern; consequently, the same pattern would be categorized differently under DSF versus DIF. Table 1 offers a comparison of classification schemes for DSF versus DIF effects. If several classifications have been given in the literature, the pattern is labeled as "varies"; if no classification has been explicitly given in the literature for the specific pattern, it is listed as unknown. Although there is some overlap between Penfield's (2008) DSF taxonomy and previous descriptors for DIF patterns, to remove confusion, the DSF taxonomy should not be used to classify DIF.

Table 1. Comparison of Classifications for example DSF/DIF Item Effects

DSF Classification	DSF/DIF effects			DIF Classification
	Step 1	Step 2	Step 3	
Constant (Pervasive)	0.6	0.6	0.6	Constant
Constant (Pervasive)	-0.6	-0.6	-0.6	Constant
Constant (Pervasive)	-0.6	-0.6	-0.8	Unknown
Divergent (Pervasive)	0.6	0.8	-0.8	Unknown
Constant (Non-Pervasive)	0.4	-0.8	-0.8	Unknown
Convergent (Non-Pervasive)	0.4	0.6	0.8	Unknown
Divergent (Non-Pervasive)	0.4	0.8	-0.8	Unknown
Constant (Non-Pervasive)	0	0.5	0	Varies
Constant (Non-Pervasive)	0.5	0	0	Low shift
Constant (Non-Pervasive)	0	0	0.5	High shift
No DSF	-0.2	0	0.2	Balanced
No DSF	0.3	0	0	Low shift
No DSF	0.3	0.3	0.3	Constant
No DSF	0.3	0.3	0	Varies

Parametric versus Nonparametric Polytomous DIF and DSF Tests

Reviews for polytomous DIF statistical procedures have been provided by Millsap & Everson (1993), Penfield & Lam (2000), and Potenza & Dorans (1995). Parametric DIF/DSF tests assume that response data for a test is generated by a particular underlying mathematical model. Once this mathematical model is assumed, inferences are subsequently made based on model fit. Due to the mathematical model that parametric methods use, item parameters are an important part of estimation procedures to verify model fit and conduct additional tests based on the selected model. For parametric methods, DIF can be tested in terms of the statistical difference between particular item parameters (Lord 1980, Raju, 1988; Schulz 2011). Thus, accurate parameter estimation is essential when testing for measurement invariance; adequate model fit as well as a large sample size are necessary for accurate parameter estimates (De Ayala, 1995; Stone, 1992; Yamamoto & Muraki, 1991; Zwinderman & Van den Wollenberg, 1990).

Nonparametric methods do not impose a particular model on the response data. Thus, there is no need for an initial mathematical model to be fit in order to test for differences between group responses. With non parametric methods, assumptions have to do with the estimator used to test for differences between groups of test takers obtaining a certain score on an item. As an example, for the null hypothesis of no difference, it is expected that, over repeated samples, the estimators will follow a particular distribution and if that assumption is violated, then the null hypothesis is rejected. Mantel-Haenszel based methods have traditionally been the most widely used in the literature (Ankenmann et al., 1999; Zwick et al., 1993; Chang et al., 1996; Wang & Su, 2004). However, within the past 15 years, additional methods have been proposed to detect DIF in polytomous items that add the benefit of providing an effect size (i.e. Liu-Agresti, Cox's β) or can be used in conjuction with DSF methods (i.e. Simultaneous Step Level Test) (Penfield, 2007; Penfield & Algina, 2003; Penfield & Gattamorta, 2009). An advantage of nonparametric DIF procedures over parametric DIF procedures is that, as mentioned, parametric procedures require a larger sample size. Based on simulation studies for Mantel-Haenzel based methods, Zwick (2012) recommended 700 total sample size with at least 200 for the reference or focal group. Sample sizes for parametric methods in the literature were usually found to be in the thousands and higher (De Ayala, 2009). Another advantage of

nonparametric methods is that there is no requirement for data to adequately fit a particular model; misspecification of the model may cause the representation of the DIF effect to be invalid and inflate Type I errors (Penfield 2007, Penfield & Algina, 2003; Fidalgo & Madeira, 2008).

Methods for detecting differential item or step functioning in polytomous items can be further classified as observed score parametric (i.e. Multinomial logistic regression), observed score non-parametric (i.e Mantel Test), latent variable parametric (i.e. IRT Likelihood Ratio Test), and latent variable non-parametric (i.e. SIBTEST) (Hidalgo & Gomez, 2006; Lord 1980, Mantel, 1963; Raju, 1988; Schulz and Fraillon, 2011; Shealy and Stout, 1993). Given some of the advantages of nonparametric methods mentioned above, this review will focus on nonparametric observed score approaches.

Numerous approaches for measuring DIF in polytomous items yield only an omnibus measure of DIF (Camilli & Congdon, 1999; Chang et al., 1996; Dorans & Scmitt, 1991; Mantel, 1963; Zwick & Thayer, 1996). Omnibus DIF estimators yield an aggregated (summated) effect and display low power with effects varying in size or magnitude across steps (Penfield, 2007; Wang & Su, 2004). DSF detection procedures provide valuable information concerning precisely which steps are exhibiting between-group measurement differences that could go undetected by DIF estimators (Miller, Chahine, & Childs, 2010; Penfield, 2007; Penfield, 2008; Gattamorta & Penfield, 2012; Gattamorta, Penfield, & Myers, 2012). Additionally, another advantage of using non-parametric DSF procedures over parametric DIF/DSF procedures is that less time is consumed for detecting DSF, since parametric procedures require a separate analysis to be run for each step of each item under investigation (Penfield, 2007; Penfield, 2008; Gattamorta & Penfield, 2012).

Both DIF and DSF procedures can utilize statistical significance tests to flag items that exhibit DIF/DSF with the null hypothesis being that the effect is not statistically significantly different from zero. It is helpful to pair significance testing with effect size values to determine practical significance, however, effect size methods for polytomous DIF detection are not as well established as those for dichotomous DIF detection (Penfield & Algina, 2003; Wang & Su, 2004; Wood, 2011).

Description of Differential Step Functioning Detection Procedures

Description of the DSF detection procedures used for this study closely follows Gattamorta and Penfield (2012). The odds ratio compares the odds of successfully advancing to the j^{th} step for reference or focal group members with the same score. Examinees are divided into score groups based on raw summed scores of a test with possible score values of $k = 1, 2, 3, \ldots, S$. The score value serves as a proxy for ability level and will be used as the matching variable between reference and focal groups. As shown in Table 2, a 2 x J x K contingency table is created for each item which represents 2 comparison groups (reference and focal) x J response categories x K score/ability levels.

A ratio of odds of success over the jth step for the reference group over the focal group is estimated using:

$$\hat{\boldsymbol{\alpha}}_j = \frac{\sum_{k=1}^{K} A_{jk} D_{jk} / N_{jk}}{\sum_{k=1}^{K} B_{jk} C_{jk} / N_{jk}} \quad (1)$$

Let A_{jk} and B_{jk} represent the number of reference group members that were successful and unsuccessful at the j^{th} step, respectively; let, C_{jk} and D_{jk} represents the number of focal group members that were successful and unsuccessful at the j^{th} step, respectively. Also, N_{jk} represents the sum of A_{jk}, B_{jk}, C_{jk} and D_{jk}. This estimator is equivalent to the Mantel-Haenszel odds ratio for dichotomous items; each step is treated as a dichotomy. A polytomous item is dichotomized differently under the adjacent category versus the cumulative category approaches such that A_{jk}, B_{jk}, C_{jk}, D_{jk} and N_{jk} are defined differently for each approach. The DSF approaches test how large the log-odds ratio effect is between cells of two groups based on assuming asymptotic normality of the log-odds ratio estimators (Penfield, 2008).

Table 2. The *k*th level of a 2 x J contingency table

Group	Item Score					Total
	y_1	y_2	y_3	...	y_J	
Reference	n_{R1k}	n_{R2k}	n_{R3k}	...	n_{RJk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	...	n_{FJk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	...	n_{+Jk}	n_{++k}

Adjacent Category Approach

Under the adjacent category approach for step 1, Ajk and Bjk represent the number of reference group members that obtained a score of 1 and 0, respectively; and, Cjk and Djk represent the number of focal group members that obtained a score of 1 and 0, respectively. Under this approach, Njk, represents the total number of examinees that scored 0 or 1; scores of 2 or higher would not be considered at the first step. Under the adjacent category approach for an item with 4 categories, a DSF effect at step 3 indicates a more difficult transition from a score of 2 to 3 for focal group versus reference group members.

Cumulative Category Approach

This approach has also been called common log odds or cumulative common log odds, but will be referred to as cumulative category log odds in this study to remain consistent. Under the cumulative approach for step 1, A_{jk} represent the number of reference group members that obtained a score of 1, 2, or 3 and B_{jk} represent the number of reference group members that obtained a score of 0; and, C_{jk} represents the number of focal group members that obtained a score of 1, 2, or 3 and D_{jk} represents the number of focal group members that obtained a score of 0. Under this approach, N_{jk} represents the total number of examinees. Under the cumulative category approach for an item with 4 categories, a DSF effect at step 3 indicates a more difficult transition from a score of 0, 1,

or 2 to a score of 3 for focal group versus reference group members.

The natural logarithm of $\hat{a}_j$ is denoted $\hat{\lambda}_j$. When $\hat{\lambda}_j$ is positive, DSF is in favor of the reference group; when $\hat{\lambda}_j$ is negative, DSF is in favor of the focal group, and a zero value indicates no DSF. Since the estimator $\hat{a}_j$, is consistent in scale and direction with Mantel-Haenszel estimator for DIF in dichotomous items, the ETS classification for effect sizes (Zeiky, 1993) can be utilized. Thus, $\left|\hat{\lambda}_j\right| < 0.43$ corresponds to a small DSF effect, $0.43 \leq \left|\hat{\lambda}_j\right| \leq 0.64$ corresponds to a moderate DSF effect and $\left|\hat{\lambda}_j\right| \geq 0.64$ corresponds to a large DSF effect (Penfield, 2007; Penfield, Alvaraz, et al., 2009). ETS determined these effect size thresholds by utilizing the Mantel-Haenszel estimator, $\hat{\alpha}_{MH}$, (equivalent to $\hat{a}_j$ as noted above) which is on a scale of 0 to infinity where values greater than 1 favor the reference group. Holland and Thayer (1988) developed a transformation of this estimator, *MH D-DIF* (Mantel Haenszel delta difference) to correspond with the ETS ability scale which is the delta scale (linear transformation of the inverse normal equivalent). MH D-DIF is defined as -2.35*ln($\hat{\alpha}_{MH}$). By utilizing the Mantel Haenszel Chi Square statistic, and setting MH D-DIF equal to 1, ETS classified a small effect as one with a non-significant *p*-value and ln($\hat{\alpha}_{MH}$) (which is equivalent to $\hat{\lambda}_j$) less than 0.43 (here we see that 1/2.35 gives us 0.43). Since values greater than 1 favor the reference group, MH D-DIF was then set to 1.5 to determine the large and significant effect of 0.64 or greater (here, 1.5/2.35 gives us 0.64). Finally, values in between these endpoints constitute a medium effect size.

The standard error of $\hat{\lambda}_j$ can be computed by (Camilli & Shepard, 1994; Penfield & Camilli, 2007):

$$SE(\hat{\boldsymbol{\lambda}}_j) = \sqrt{\frac{\sum_{k=1}^{K} T_k^{-2}(A_{jk}D_{jk} + \hat{\boldsymbol{\alpha}}_j B_{jk}C_{jk})(A_{jk} + D_{jk} + \hat{\boldsymbol{\alpha}}_j B_{jk} + \hat{\boldsymbol{\alpha}}_j C_{jk})}{2\left(\sum_{k=1}^{K} \frac{A_{jk}D_{jk}}{T_k}\right)^2}} \qquad (2)$$

In equation (2) T_k is the total number of individuals at the *k*th stratum of ability. The test

statistic, is distributed as standard normal under the hypothesis of no DSF:

$$z(\hat{\boldsymbol{\lambda}}_{+}) = \frac{\hat{\boldsymbol{\lambda}}_{+}}{SE(\hat{\boldsymbol{\lambda}}_{+})}, \tag{3}$$

A disadvantage of DSF estimators is the lower level of power due to smaller sample sizes within each item category. In this case, the effect size may be more informative than the significance test, although both should still be used (Gattormata, et al., 2012). Statistical assumptions for DSF methods are similar to those for the MH statistic for dichotomized response data (Mantel & Haenszel, 1959). First, each observation comes from a different subject, subject groups are randomly selected and no subjects are purposely omitted. Secondly, all observations are identically distributed which means all observations are obtained in the same way. Thirdly, under the null hypothesis of no partial association and the assumption of fixed marginal totals, the underlying probability model for the observations is hypergeometic. Finally, the proportional odds function is consistent across *j* steps. *R* functions for the adjacent category log odds ratio (AC-LOR) and cumulative category log odds ratio (CU-LOR) are presented in Appendices B1 and B2.

Comparison to Other DSF Methods

Other non-parametric methods that give an index of DSF include the continuous ratio estimator (CR-LOR), SIBTEST, standardized p-difference index, and the Yanagawa-Fuji estimator (Yanagawa & Fujii, 1990; Zwick, Donoghue, and Grima, 1993). The continuous ratio (CR-LOR) estimator compares the number of reference/focal group responses at step *j* to the number at *j*+1, *j*+2, ..., *J* steps; this estimator is compatible with the continuous ratio model (CRM) which specifies an *a* parameter for each step (Hemker, Van der Ark, Sijtsma, 2001). The CRM is not as widely accepted as existing IRT models (i.e GRM, PCM); although Penfield (2008) offered a parameterization for CR-LOR that made it more consistent with IRT models, this estimator is rarely mentioned in the DIF literature.

The standardization method proposed by Dorans and Kulick (1986) and simultaneous bias test (SIBTEST) proposed by (Shealy & Stout, 1993) utilize a weighted

difference between the focal group's average score on an item and the reference group's average score on an item. Both methods provide an effect size. An advantage that the common odds ratio approach has over these non-parametric approaches is that the same general criteria and guidelines can be used for interpreting omnibus DIF and DSF effects since both DSF and DIF common odds ratio approaches share approximately the same metric (Penfield, 2007). Additionally, the standardization and SIBTEST approaches are mean-difference indices that are dependent on the number of response options under investigation when interpreting the effect.

Yanagawa and Fujii (1990) developed a conditional test for homogeneity of the odds ratios of I x J x K contingency tables based on a multiple hypergeometric distribution. The test included a correction in order to yield an asymptotic chi-squared distribution and an algorithm that uses the generalized Mantel-Haenszel estimator to calculate the test. However, the Yanagawa-Fujii estimator does not yield an index of measurement equivalence that is theoretically consistent with the step functions underlying polytomous IRT models (i.e. Partial Credit or Graded Response Models) (Penfield, 2007; Wang & Su, 2004). Thereby, it is argued that the interpretation for this estimator's effect is confounded.

Description of Polytomous DIF Detection Procedures

Mantel Test

First used in epidemiological research, the popularity of the MH statistic in the 1980s, and even now, for detecting DIF in dichotomous items led to the implementation and use of the Mantel test for polytomous items since the 1990s (Dorans & Holland, 1993; Holland & Thayer, 1988; Penfield, 2001; Welch & Hoover, 1993; Zwick et al, 1993; Woods, 2011). Mantel (1963) extended the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) statistic for 2 x 2 contingency tables so that 2 x *k* Chi Square tables, each with a single degree of freedom, would be calculated for the Mantel test. To implement the Mantel test for the reference and focal group, items are organized into 2 x J x K contingency tables where J is the number of response categories or score levels for a polytomous item. The variable K corresponds to the number of score levels for the matching variable. Thus, there will be a 2 x J contingency table at each level of K. An

example is given in Table 2. The values $y_{1,}y_{2,\ldots,}y_{J}$ correspond to the J scores of the item. The values n_{RK} and n_{FK} correspond to the number of respondents in the reference and focal group, respectively, who received a score of y_J at level K. The symbol + denotes summation over a particular index. To create the test statistic for the Mantel test, the sum of scores of the focal group at the *k*th level of the matching variable is calculated by

$$F_k = \sum_{j=1}^{J} y_j n_{Fjk} \tag{4}$$

Under the null hypothesis of no difference between item means of the reference and focal group (column and rows are independent) the expected value and variance of F_k are given as:

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_{j=1}^{J} y_j n_{+jk}, \tag{5}$$

and

$$Var\,(F_k) = \frac{n_{R+k} n_{F+k}}{n_{++}^2 (n_{++k} - 1)} \left\{ \left(n_{++k} \sum_{j=1}^{J} y_j^2 n_{+jk} \right) - \left(\sum_{j=1}^{J} y_j n_{+jk} \right)^2 \right\}, \tag{6}$$

The test statistic of the Mantel is

$$\chi^2_{Mantel} \frac{\left(\sum_{k=1}^{K} F_k - \sum_{k=1}^{K} E(F_k) \right)^2}{\sum_{k=1}^{K} Var\,(F_k)} \tag{7}$$

Under the null hypothesis, the test statistic in equation (7) has a chi-square distribution with 1 degree of freedom. If the null hypothesis is rejected, item means for the reference and focal group differ for members who are at the same score level *k* and the item is identified as exhibiting DIF. Assumptions for generalized forms of MH statistic (i.e. Mantel test, Generalized Mantel-Haenszel) follow those listed in the DSF section (Somes, 1986). The main differences in the generalized forms of the MH statistic are how the response data is defined (categorical/nominal, rank/ordinal, or interval-scale). Additionally, the underlying probability model for the observations is multivariate

hypergeometic and the proportional odds function is not required to be the same across all *k*. An *R* function for the Mantel Test is presented in Appendix B3.

Generalized Mantel-Haenszel

The Generalized Mantel-Haenszel (GMH) statistic was introduced as a multivariate generalization of the Mantel-Haenszel procedure (Mantel & Haenszel, 1959; Somes, 1986). Unlike the Mantel Test, the GMH assumes the data are nominal and has been used as another method for detecting polytomous DIF (Atbar, 2007; Zwick et al., 1993). The test statistic is given by

$$\chi^2_{GMH} = \left[\sum \mathbf{R}_k - \sum_{k=1}^{K} E(\mathbf{R}_k)\right]' \left[\sum \mathbf{V}(\mathbf{R}_k)\right]^{-1} \left[\sum \mathbf{R}_k - \sum E(\mathbf{R}_k)\right] \tag{8}$$

where R_k is a 1 x (M-1) vector of the frequencies of the reference group for item score M -1 categories at the k^{th} level of the matching variable. This is denoted

$$\mathbf{R}_k' = \left[n_{1rk} n_{2rk} \ldots n_{(M-1)rk}\right] \tag{9}$$

$E(\mathbf{R}_k')$ is a 1 x (M-1) vector:

$$E(\mathbf{R}_k') = \frac{N_{rk}}{N_k} T_k', \tag{10}$$

for this vector, T_k' is a 1 x (M - 1) vector of the frequencies in both the reference and focal groups for M -1 item score category at the k^{th} level of the matching variable,

$$\mathbf{T}_k' = \left[n_{1k} n_{2k} \ldots n_{(M-1)k}\right] \tag{11}$$

$$Var(\mathbf{R}_k') = \frac{N_{rk} N_{fk}}{N_k^2 (N_k - 1)} \left[N_k diag\mathbf{T}_k - \mathbf{T}_k \mathbf{T}_k'\right], \tag{12}$$

Under the null hypothesis of no general conditional association between reference and focal group response data, (8) follows approximately a chi-square distribution with (rows -1) x (columns - 1) degrees of freedom (Fidalgo, 2008). When the null hypothesis is

rejected, the distribution of the response variable differs in non-specific patterns across levels of the matching variable *K*. An *R* function for the Generalized Mantel-Haenszel procedure is presented in Appendix B4.

Liu-Agresti

Liu and Agresti (1996) provided an estimator of the common odds ratio as a generalization of the $\hat{\alpha}_{MH}$ for all *K* levels of ordinal response variables. The Liu-Agresti method was proposed as a method for detecting DIF in polytomous items (Penfield & Algina, 2003; Penfield, 2007); although it has not been as widely used as other methods, such as Mantel-Haenszel based procedures. A benefit of the Liu-Agresti estimator is that, along with a statistical test, it also provides a measure of effect size for the magnitude of DIF that uses the same classification scheme as Mantel-Haenszel estimator for dichotomous items. To calculate the Liu-Agresti estimator, one would simply sum equation (1) over all levels of the matching variable *K:*

$$\psi_{LA} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J} A_{jk} D_{jk} / N_{jk}}{\sum_{k=1}^{K}\sum_{j=1}^{J} B_{jk} C_{jk} / N_{jk}}. \tag{14}$$

To produce an estimator having the same scale as $\hat{\alpha}_{MH}$, one would use

$$\hat{\alpha}_{LA} = \frac{1}{\hat{\psi}_{LA}}. \tag{15}$$

Similar to $\hat{\alpha}_{MH}$, the natural log, ln($\hat{\alpha}_{LA}$) would be taken to produce a symmetric scale so that ln($\hat{\alpha}_{LA}$) = 0 means no DIF, ln($\hat{\alpha}_{LA}$) > 0 favors the reference group and ln($\hat{\alpha}_{LA}$) < 0 favors the focal group. The variance is given by

$$\text{Var}[\ln(\hat{\alpha})] = \frac{\sum_{k=1}^{K} \hat{\xi}_k}{\hat{\alpha}_{+}\left[\sum_{k=1}^{K}\sum_{j=1}^{J} B_{jk} C_{jk} / N_k\right]^2}, . \tag{16}$$

Where

$$\hat{\xi}_k = \sum_{j=j'=1}^{J} \hat{\varphi}_{jj'k} + 2\sum_{j<j'}^{J} \hat{\varphi}_{jj'k} \ .$$

and

$$\hat{\varphi}_{jj'k} = \frac{N_{Rk}N_{Fk}}{N_k^2}\left\{\frac{\hat{\alpha}_+ B_{jk}C_{j'k}}{N_{Rk}}\left[1+(\hat{\alpha}_+ -1)\frac{C_{j'k}}{N_{Fk}}\right]\right.$$

$$\left.+\frac{A_{jk}D_{j'k}}{N_{Fk}}\left[\hat{\alpha}_+ -(\hat{\alpha}_+ -1)\frac{A_{jk}}{N_{Rk}}\right]\right\}, j \le j' = 1,...,J. .$$

The test statistic used to test the null hypothesis of no DIF is

$$z(\ln(\hat{\alpha}_+)) = \frac{\ln(\hat{\alpha}_+)}{SE(\ln(\hat{\alpha}))}. \qquad (17)$$

An underlying assumption for the Liu-Agresti estimator is that response data follows a proportional odds structure. When a proportional odds ratio holds, the relationship between the cumulative logit to the explanatory variables is linear and also the same for all *j* categories (Penfield & Algina, 2003). An *R* function for the Liu-Agresti procedure is presented in Appendix B5[1].

Simultaneous Step Level Test

A procedure that pursues a simultaneous test of no DSF across *J* steps of the studied item was introduced by Penfield (2007; 2009). The simultaneous step level test (SSL) is a non-traditional way of detecting DIF in polytomous items that appears only in the DSF literature. It utilizes CU-LOR to test for DSF at each step of the item. The condition of no DIF is satisfied if no DSF exists at each of the *J* steps and thus the null hypothesis of no DIF can be rejected if one or more of the *J* null hypotheses of no DSF is rejected using

[1] The standard error for this procedure was coded with help from Dr. Randall Penfield.

the $z(\hat{\boldsymbol{\lambda}}_{+})$ test statistic from equation (3). Penfield (2007) suggested using the Bonferroni adjustment to adjust the Type I error rate, α, of each test of DSF within an item to α/*J*. Underlying statistical assumptions for the SSL test are similar as those listed in the DSF section since it depends on the CU-LOR statistic. The z-test of DSF at each step *j* is based on the assumption of asymptotic normality of the CU-LOR statistic (Penfield, 2007). An *R* function for the SSL test is presented in Appendix B6.

Comparison to Other Observed Score Approach DIF Methods

The standardization mean difference (SMD) approach (Dorans & Scmitt, 1991) utilizes empirical item test regressions, comparing the difference at each score level between weighted frequencies of reference and focal group's members at each score level (Potenza and Dorans, 1995). A correction is used to adjust for differences in distributions of the reference and focal groups across the matching variable. The dependent variable, item score, has $j = 1, 2, ..., J$ levels or categories. The differences are weighted by the frequencies of focal group members with item score *j* at matching variable level *k*. Weighted differences are then summed across the matching variable to determine the measure of DIF. The null DIF definition can be expressed as zero difference in expected item score given the matching variable or no difference in item-test regressions between the reference and focal groups. Large values of |SMD| indicate DIF; however, Dorans & Schmitt (1991) did not provide clear guidelines regarding what would be considered "large". Zwick & Thayer (1996) proposed two estimations of the standard error for SMD based on distributional assumptions about the data. Both the SMD and Mantel test utilize expected/average item scores and have associated statistical tests. However, while Mantel has one statistical test associated with it, SMD has two possible statistical tests; also, the SMD estimator can provide information on the magnitude of DIF, but there are no clear guidelines to classify the size of these magnitudes. Regarding comparisons to GMH, the GMH procedure compares entire item response distributions between reference and focal groups, conditioned on the matching variable, while the Mantel test and SMD procedure are sensitive to the mean differences of these distributions. No comparison information is available for the SMD procedure and the Liu-Agresti estimator or SSL test.

The HW1 and HW3 statistics both use two sample *t*-test statistics to detect departure from null DIF for polytomously scored items (Welch & Hoover, 1993; Potenza & Dorans, 1995; Wood, 2011). For HW1, the difference in item score at each level of the matching variable between reference and focal groups is computed and converted to a *t*-statistic by dividing by the pooled standard error estimate. These *t*-statistics are summed across the matching variable and divided by the square root of the independent *t*-statistic variances; this final *t*-statistic is normally distributed with mean 0 and standard deviation 1. On the contrary, the HW3 statistic weights each test statistic by the reciprocal of the sampling variance. A correction is provided at each level of the matching variable to account for small sample size; the final statistic is also normally distributed with mean 0 and standard deviation 1. Thus, the HW1 and HW3 methods utilize mean differences similar to the standardization and Mantel methods, but, like the Mantel test, HW1 and HW3 are statistical tests. There are few studies comparing HW1 and HW3 to other methods, however, Wood (2011) discovered that the Liu-Agresti estimator and Mantel Test provided better controlled Type 1 error rates when focal group size was as small as 40 (reference group size was 400).

Finally, Cox's β, can be simply computed as the square root of the Mantel Test statistic (Cox, 1958). However theoretically, the Cox's β procedure assumes the data follows a non-central hypergeometric distribution, while the Mantel test assumes a central hypergeometric distribution. An estimate of $\hat{\beta}$ and its variance was given by Camilli & Congdon (1999) which produced a *Z* statistic that is normally distributed under the null hypothesis of no DIF. Large values of |Z| indicate DIF; additionally, since $\hat{\beta}$ is approximately normally distributed and symmetric about zero, it can be used as an effect size estimator for the magnitude of DIF. Although, not frequently used in the literature, Penfield & Algina (2003) found that, statistically, Cox's β performed identically to the Liu-Agresti estimator in terms of Type I error and power rates (which in turn, assumes that the Liu-Agresti estimator would perform statisically identically to the Mantel test, as well). Although both the Liu-Agresti and Cox's β procedures have an advantage of being also utilized as effect size measures, the Liu-Agresti effect size estimator can be interpreted using the same classification scheme as the Mantel-Haenszel estimator for

dichotomous items as presented by ETS (Penfield, 2003). Comparisons of Cox's β to the GMH method and SSL test were not found in the literature

Addressing Inflated Type I errors in DIF/DSF Test for Polytomous Items

Several factors may cause Type I error when detecting DSF/DIF, including differences in reference and focal group distributions (i.e. impact), multiple significance testing of items or item steps, and multiple items with DIF/DSF effects (Fleming, Ross, Tollefson, Green 1998; Gilmore 2014; Penfield, 2008; Taylor & Lee, 2012; Wang & Su, 2004). Purification has been used as a way for reducing bias in the matching variable and subsequently Type I error; however, purification was found to be more effective for dichotomous versus polytomous Mantel-Haenszel methods and, in practice, it may also diminish the quality of the construct being measured (Taylor & Lee, 2012; Wang & Su, 2004). Statistical adjustments for Type I error can be investigated as another solution for reducing bias in the matching variable.

Prior to Kim (2010), who investigated the Type I error rate of several DIF procedures and three adjustments to the procedures - Bonferonni (1936), Benjamini-Hochberg (BH; 1995), and Holm (1979) - only one DIF study was found that incorporated a Type I error adjustment. Steinberg (2001) used the BH method to adjust for Type I errors when investigating if threshold differences/bias would occur if an item was taken independently versus paired with another item. Other studies used adjustments for paired comparisons, not DIF, and the methods used were either the Bonferroni or Benjamini- Hochberg method (Kaya, Leite, Miller, 2015; Kim, 2010; Steinberg 2001; Thissen, Steinberg, Kuang, 2002; Williams, Jones, Tukey, 1999). After Kim (2010), two other DIF studies used either a Bonferroni or BH adjustment when trying to identify DIF (DIF methods included MH, Logistic Regression, and SIBTEST) (Kabasakal, Arsan, Gok, Kelecioglu, 2014; Gilmore, 2014). Based on these studies, both non and parametric methods can benefit from adjustments in the case of large sample sizes and long tests, but non-parametric can benefit even for short tests. Penfield (2007) and Penfield, Alvarez, and Lee (2008) consistently used the Bonferonni correction for the SSL method. It is deemed necessary because the SSL method tests each step within an item for DSF to determine if the item has DIF. Besides Penfield (2007) and Penfield, Alvarez, and Lee

(2008), it was difficult to find other studies that utilized statistical adjustments for inflated Type I error when investigating DIF/DSF in polytomous items. The next few paragraphs will describe the three adjustment methods that have been found in the DIF/DSF literature.

Dunn-Bonferroni

The Dunn-Bonferroni method, typically referred to as the Bonferroni method (Bonferroni, 1936; Miller 1981) is commonly used in DSF studies; though it tends to be conservative, it is the simplest method for adjusting possible inflated Type I error rates due to multiple significance tests (Penfield 2007, 2008). To find the Bonferroni adjusted Type I error rate, the nominal significance level (i.e. $\alpha = .05$) will be divided by the number of significance tests being conducted.

Benjamini and Hochberg

The Bonferroni method, however, has been shown to be unnecessarily stringent for many practical situations; therefore, a more recently developed method to correct for multiplicity, the Benjamini and Hochberg (BH) method, has been recommended by What Works ClearingHouse (2014). As mentioned, the BH method is supposed to be less conservative than the Bonferroni method (Kim, 2010; William, Jones, & Tukey, 1999). The BH method defines a sequential *p* – value procedure as follows:

1. Rank order the *p* – value for each item from smallest to largest, $P_1,P_2,...,P_J$
2. Retain the largest *p* – value, P_J.
3. Multiply the second largest *p* – value, P_{J-1}, by the number of items in the test and divide by its rank ($P'_{J-1} = P_{J-1} * n/(n\text{-}1)$ where *n* is the total number of items).
4. Choose for the value of P'_{J-1} the minimum value of P'_{J-1}, P_J, or 1 (ensures *p*-values remain monotonically decreasing); if $P'_{J-1} < 0.05$, it is significant.
5. Multiply the third largest *p* – value, P_{J-2}, by the number of items and divide by its rank ($P'_{J-2} = P_{J-2} * n/(n\text{-}2)$ where *n* is the total number of items).
6. Choose for the value of P'_{J-2} the minimum value of P'_{J-2}, P_{J-1}, or 1; if $P'_{J-2} < 0.05$, it is significant.

7. In general, $P'_{J-i} = P_{J-i}* n/(n\text{-i})$ and choose for the value of P'_{J-i} the minimum value of P'_{J-i}, P_{J-i-1}, or 1. If $P'_{J-i} < 0.05$, it is significant. Continue the procedure until the smallest p – value is corrected.

Here is an example of the procedure with given p-values 0.105, 0.02, 0.088:

1. 0.105 0.088, 0.020
2. $P_J = 0.105$
3. $P'_{J-1} = 0.088*3/2 = 0.132$
4. $P'_{J-1} = \min(.132,.105,1) = 0.105$ (not significant)
5. $P'_{J-2} = 0.020 * 3/1 = 0.060$
6. $P'_{J-2} = \min(0.105, 0.060,1) = 0.060$ (not significant)
7. Now we have all corrected p-values, P_J, P'_{J-1}, P'_{J-2}: 0.105, 0.105, 0.060.

Holm

The Holm's procedure is also said to be less conservative than Bonferroni's method; it is implemented in two stages. In the first stage, the total number of p values will be ordered from smallest to largest. If $i*$ is the smallest integer from 1 to k such that $p_{(i*)} > \alpha/(k - i* + 1)$, then all hypothesis tests corresponding to the integer values before i* will be rejected and all remaining hypothesis tests from $i*$ to k will be retained. If no such integer meets the criteria for $p_{(i*)}$, then all k hypothesis tests will be rejected. This is illustrated using the same p-values 0.105, 0.020, 0.088:

1. $p_{(1)} = 0.020$, $p_{(2)} = 0.088$, $p_{(3)} = 0.105$
2. For $p_{(1)}$ we have $\alpha/(k - i* + 1) = 0.05/(3\text{-}1\text{+}1)$. So $p_{(1)} > 0.017$
3. For $p_{(2)}$ we have $0.05/(3\text{-}2\text{+}1)$. So $p_{(2)} > 0.025$
4. For $p_{(3)}$ we have $0.05/(3\text{-}3\text{+}1)$. So $p_{(3)} > 0.05$

Since $i* = 1$ is the smallest integer that fits the criteria, none of the p-values are rejected, they are all retained. This is more obvious from looking at the adjusted p-values and comparing each one to the criteria $\alpha = 0.05$. To calculate the adjusted p-values, we use $p'_{(i*)} = p_{(i*)}(k - i* + 1)$, such that:

$p'_{(1)} = 0.020(3\text{-}1\text{+}1) = 0.060$

$p'_{(2)} = 0.088(3\text{-}2\text{+}1) = 0.176$

$p'_{(3)} = 0.105(3\text{-}3\text{+}1) = 0.105$

Thus, none of the adjusted *p*-values are less than 0.05.

Other Methods for Addressing Inflated Type I Error Rates

There are several variations to the adjustments mentioned in this section. However, they will only be mentioned very briefly because they are not found in DIF literature and will not be used in this study. The Benjamini-Yakutieli procedure (Benjamini & Yekutieli, 2001) is similar to the Benjamini-Hochberg procedure except that it allows for dependence among hypothesis tests. Both the Benjamini-Hochberg and Benjamini-Yekutieli procedures control the false discovery rate, the expected proportion of false discoveries among the rejected hypothesis (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). The false discovery rate is less restrictive than the family-wise error rate, which make these methods more powerful than Holm and Bonferonni (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). Two methods that are valid when hypothesis tests are independent or non-negatively associated are Hochberg (1988) and Hommel (1988) procedures. Although Hochberg's p-values can be computed more quickly, Hommel's procedure is more powerful.

There are also non-statistical ways of addressing Type I error rates that were found in the literature. These include manipulating sample size, test length, magnitude of DIF, percentage of items with DIF, or incorporating purification procedures (Fidalgo, Mellenbergh, & Muniz, 2000; Guilera, Gomez-Benito, Hidalgo, Sanchez-Meca, 2013; Holland & Thayer, 1988; Kim, 2010; Narayanan & Swaminathan, 1994; Gilmore, 2014). Although the alpha level, α, could be decreased to decrease Type I error, it was difficult to find studies with an α level less than .05. Only a few studies have reported more than one significance level (Ankenmann et al., 1999; Cohen, Kim, Wollack,1996; Kim, Cohen, Kim, 1994; Narayanan & Swaminathan, 1996), although Fidalgo, Ferreres, Muniz (2004) suggested increasing the significance level to .20 to increase power.

Typically sample or test features are utilized in DIF studies to control for Type I error (Kim, 2010; Wang & Su, 2004). Both reducing sample size or shortening tests can reduce false positives when attempting to detect DIF (Kim, 2010; Wang & Su, 2004); although percentage of items with DIF and average signed area may have more of an impact (Narayanan & Swaminathan, 1994; Wang & Su, 2004).

Purification procedures for reducing inflated Type I errors are commonly used in simulation studies regarding DIF (Clauser, Mazor, & Hambelton, 1993; Fidalgo, Mellenbergh, & Muniz, 2000; Holland & Thayer, 1988; Kwak, Davenport, Davison, 1998; Navas-Ara & Gomex-Benito, Wang & Su, 2004). Iterative procedures are particularly effective, although Wang & Su (2004) found it more beneficial in dichotomous MH procedures versus Mantel procedures for polytomous items. In practical situations, as mentioned, purification may also diminish the quality of the construct being measured (Taylor & Lee, 2012).

Impact also causes Type I inflation in simulation studies (Bolt & Gierl, 2006; Clauser, Mazor, & Hambelton, 1993; Wang & Su, 2004). Although impact cannot be directly manipulated in practice, if the reference group differs substantially in ability, (typically 1 standard deviation or greater) from the focal group, it may be unwise to conduct a DIF analysis since differences in scores may be due to differences in ability and not item bias (Dorans & Holland, 1993; Meulders & Xie, 2004; Wang & Su, 2004).

Factors Considered in DIF/DSF Simulation Studies

There are common factors considered in simulations studies for detecting DIF/DSF that have been shown to affect Type I error and power rates for polytomous items. These factors include: Sample characteristics (sample size, sample size ratio, impact), test/item characteristics (test length, number of item score levels), and analysis characteristics (generating model and parameters, effect size, percent of DIF, DIF/DSF pattern, purification procedure, inclusion of studied item(s)). For this section, only a few studies (less than 5) were found that utilized both DIF and DSF in their investigation.

Sample Characteristics

In DIF studies, sample sizes considered have been as large as 6,000, with 3,000 each for the reference and focal groups, (Bolt & Gierl, 2006; Shealy & Stout, 1991; Zwick & Thayer, 2002) or as small as 100, with 50 each for the reference and focal groups (Fidalgo et al., 2007; Muniz et al., 2001). For DSF studies, in particular sample sizes were usually 1,000, with 500 each for the reference and focal groups (Gattamorta & Penfield, 2012; Penfield, 2007, Penfield, Alvarz, & Lee, 2008, Penfield, 2010). Studies utilizing IRT models used larger sample sizes than studies utilizing non-parametric

observed score methods. Regardless of DIF/DSF method, typically larger sample sizes produces higher power, but also higher Type I error rates.

Sample size ratios ranged from 1:1 to 10:1 in DIF studies; generally, unequal sample size ratios were not used in DSF simulation studies (Atar, 2007; Penfield, 2007, Penfield & Algina, 2003). A large ratio, for instance, may represent the difference in size between a majority and minority group; while equal sizes may represent a group of males versus females. As a reminder, there are only a handful of DSF studies when compared to DIF studies. The main reason given for equal sample size ratios in DSF simulation studies was to keep the analysis manageable. Due to the number of conditions, variations in sample size were left for future work even though sample sizes were very unequal in applied studies (ranging from 2:1 to 40:1). Regardless of DIF method, power to detect uniform DIF tends to decrease as the ratio of reference to focal group members increases; also, Type I error tends to be higher when sample sizes are equal between reference and focal groups (Wang & Su, 2004; Zwick, 1996). No information is available regarding the effect of sample size ratio on DSF methods.

When no impact exists, the mean difference between ability distributions of the reference and focal groups is zero. To produce impact, studies generally have the mean of the two distributions differ by as large as 1.5 standard deviations; differences between mean distribution values were found to be 0, 0.5, 0.75, 1 and 1.5 standard deviations (Penfield, 2008; Raju, Fortmann-Johnson, Kim, Morris, Nering, Oshima, 2009; Su & Wang, 2005; Woods, 2009). Type I error rates tend to become inflated as impact becomes large and worsens with model misfit.

Test Characteristics

When test length is used as a factor in simulation studies, number of items have been 10, 20, 25, 30, 40, 50, and as high as 80; tests with polytomous items are generally shorter than with dichotomous items. When comparing the Mantel test and GMH methods, Wang & Su (2004) found that under test lengths of 10, 20, and 30, as impact increased to 1.5 under the PCM model, the Type I error and power were identical for Mantel and GMH methods.. Zwick et al. 1993 also investigated different test lengths comparing Mantel and GMH procedures; it is assumed that performance between the two

methods were similar because the authors did not report results suggesting otherwise. Kim (2010) recommended that Type I error in non-parametric tests (MH versus logistic regression and SIBTEST) could be improved for shorter tests (20 versus 40) by using statistical adjustments to control for Type I error. Test lengths were not manipulated in studies found using Liu-Agresti, SSL, and DIF methods. Generally test lengths for these methods were fixed between 20 and 30 items.

Although it has been suggested that studies investigate the effect of the number of item score levels on DIF/DSF methods (Atar, 2007; Penfield, 2007, Penfield & Algina, 2003), studies typically focus on a particular number of score levels for all items. Otherwise, investigations that included mixed format tests (i.e. 20 dichotomous items and 4 polytomous items) only performed the DIF or DSF analysis on the polytomous items. Atar (2007) did investigate Type I error and power of parametric based DIF methods (IRT likelihood ratio test, logistic regression, generalized linear latent and mixed models) on dichotomous, polytomous, and mixed format tests and found that these particular methods performed similarly across tests formats. Differences were caused by sample size and DIF magnitude variations across test formats. Penfield & Algina (2003) recommended investigating the effect of the number of categories on the determination of the effect size for the Liu-Agresti method and Penfield (2007) also suggested using the number of score levels as a factor when determining the performance of the DSF estimators.

Analysis Characteristics

In DIF and DSF studies, the GRM and PCM IRT models are commonly used to generate simulated response data; at times, the Generalized PCM (GPCM) has been used, which includes the possibility of item discrimination for the PCM model. Penfield (2008) used the continuation ratio model (CRM) in one DSF study, but this model is not widely used in the literature. Results have indicated that power may be higher as the values of the discrimination parameter, *a,* increase, however Type I error becomes inflated, particularly in the presence of impact and constant DIF; on the contrary, as the difficulty parameter, *b*, increases, Type I error rates appear to become more controlled (Jing & Stout, 1998; Monahan & Ankenmann, 2005; Penfield & Algina, 2003). Additionally,

Mantel test and GMH methods exhibit inflated Type 1 error rates when the model used to generate response data is an item response model that does not follow the family of Rasch models and where latent trait distributions of the reference and focal group differ substantially (Chang & Mazzeo, 1994, Chang, Mazzeo, and Roussos, 1996; Meredith & Millsap, 1992; Roussos & Stout, 1996, Zwick, 1990, Wang & Su, 2004).

There are two approaches found in the literature to simulate effect size for DIF/DSF. The most common way is to shift the IRF to the left or right for the focal group by the desired effect size (Penfield, 2007; Penfield, 2008; Su & Wang, 2005). The second approach, particularly for non-uniform DIF is to specify and calculate a particular area under the IRF between the reference and focal groups (Finch & French 2008; Hidalgo & Gomez, 2006; Swaminathan & Rogers, 1990). For either approach, the greater the effect size, the more power DIF/DSF methods will have to detect the effect.

Percent of items with DIF in simulation studies are typically around 10% or 20% but have been as high as 60%, 80%, and 100% in some (rare) cases (Park & Lautenschlager, Su & Wang, 2005, Woods, 2009). When the percent of items with DIF are around 10% or 20%, for most studies, this equates to 1 studied item. For percentages greater than 20% of items with DIF, purification may be needed due to the increased chance of non DIF items being flagged (Type I error) (Cohen & Kim, 1993; Hidalgo-Montesinos & Lopez-Pina, 2002). Some studies opted for a statistical adjustment instead of purification, but not for polytomous items (Penfield, 2001; Kim, 2010; Gilmore, 2014). Penfield (2001) compared GMH to MH and Bonferonni adjusted MH methods for dichotomous items with polytomous groups (more than 2) and a sample size of 1000 respondents (500 per reference and focal group). When the percent of items with DIF increased from 3% to 17%, Type I error for the GMH method ranged from roughly .04 to .11 while the rates for MH ranged from .09 to .18; the Bonferonni adjustment yielded the most conservative values. However, Wang and Su (2004) argued that average signed area (ASA), was more important than percent of items with DIF in regard to controlling Type I error, particularly for Mantel and GMH methods. ASA measures the area or the difference between the item response functions of the reference and focal group over all item score levels of the test. For polytomous items, it is found by simply adding all of the signed values of DIF and dividing by the total number of item score levels in the test. In general, ASA reflects

the degree to which the test favors the reference group over the focal group; when positive, it favors the reference group, when negative it favors the focal group and when zero it favors neither group. Typically, Mantel and GMH methods lose control over Type I error for values of ASA greater than 0.03. (Wang and Su, 2004).

As mentioned in the relevant terminology section of this review, DIF patterns that have been investigated for dichotomous items are simply non-uniform, uniform, and crossing. Although patterns for polytomous items can be more complex, the most commonly investigated patterns have been low shift, high shift, balanced and constant. Non-parametric observed score methods, including Mantel, SIBTEST, SMD, and GMH display power of at least .60 or better for constant DIF when the effect size of DIF for the item is .75; however power has been reported between .10 and .19. when the effect size is as low as .30 for an item with constant DIF (Chang et al., 1996; Zwick et al., 1993). Similarly, low or high shift DIF can be accurately detected if the effect size is large enough. Rejection rates for Liu-Agresti and Cox's β are similar to the methods mentioned, although Penfield & Algina (2003) showed that power was about .66 for a DIF effect size of .6; otherwise for DIF effect sizes less than .50, power was less than .30 when the item discrimination was greater than 1. Among the non-parametric observed scores mentioned, GMH is able to detect balanced DIF the best although still slightly low (about 0.40 on average) (Wang & Su, 2004; Zwick et al., 1993). It is unclear how the Liu-Agresti method performs with balanced DIF since studies were not found that directly used the balanced pattern with this method. The CU-OR method seems to be more affected by effect size than by pattern of DIF/DSF; values greater than .50 are needed for power to be near .80 (Penfield, 2007). Studies showing power rates for AC-OR under different patterns were not found.

When the matching variable, for instance, the total test score, used for the DIF analysis contains too many items with DIF, this total score may be a biased estimate of true ability. For the MH method with dichotomous items, studies have shown that two-stage and iterative purification can reduce bias in the matching variable and yield appropriate results. Clauser et al., 1993 showed that power was increased by as much as 50% for the MH method when percent of items with DIF was 20% and no impact was present. However, when tests are short (i.e. 10 items) and percentage of DIF items are

high (i.e. greater than 20%), purification procedures lose their efficiency (Clauser et al., 1993; Donoghue, Holland, & Thayer, 1993; Fidalgo, et al., 2000). Typically, one studied item is used in DIF analyses, removing the need for purification; a meta-analysis of over 1,800 studies using the MH method revealed that only about 30% used a purification procedure (Guilera, et al., 2013). Taylor & Lee (2012) was cautious about using purification in practice, suggesting that, although commonly used in simulation studies, removing items from the total score may impact the validity of the matching test and is not appropriate if the construct is multidimensional (Nandakumar, 1994). If purification is necessary, the two-stage procedure is recommended for the Mantel and GMH methods, although it is not as effective as when used for the MH method (Wang & Su, 2004). No information was found for purification procedures with Liu-Agresti or SSL methods. Additionally, simulation studies utilizing DSF methods did not employ purification procedures.

Zwick (1990) showed that including the studied item exhibiting DIF improved the behavior of the MH method even when data did not follow the Rasch model or the reference and focal group had differing ability distributions. Zwick et al. (1993) extended this finding to the Mantel and GMH procedures for polytomous items. Even though these studies specifically addressed Mantel Haenszel based methods, more recent studies utilizing other methods for DIF and DSF continue to include the studied item within the matching variable; otherwise, if more than one item is being studied, a purification procedure is usually implemented when investigating power and Type I error (Atar, 2007; Kaya, Liete & Miller, 2015; Kim, 2010; Penfield, 2007; Penfield, 2008). However, Wang & Su (2004) argued that more than one studied item can be included in DIF analysis as long as the average signed area between reference and focal group response function remain below about 0.03 for Mantel based methods.

DIF/DSF Detection in Applied Research

This section will focus on applied studies that utilized DSF or both DIF and DSF detection methods for polytomous items. Studies found were Gattamorta & Penfield (2012), Gattamorta, Penfield, & Meyers (2012), Miller et al., 2010, and Penfield (2008). Although the number of studies found were few, this section is intended to give insight on

test and item characteristics, sample sizes and sample size ratios, as well as methods used.

Types of tests that were utilized in these studies include a grade 5 and grade 8 statewide Science assessment (Gattamorta & Penfield, 2012), large-scale administered Canadian achievement test (School Achievement Indicators Program--SAIP) (Gattamorta, Penfield, & Meyers, 2012), grade 9 Assessment of Mathematics and a corresponding teacher questionnaire (Miller et al., 2010), and a test that combined items from the National Assessment of Educational Progress (NAEP), Third International Mathematics and Science Study (TIMSS) and the Florida Comprehensive Assessment Test (FCAT) (Penfield, 2008). Test lengths ranged from 16 to 36 items, where half the studies used a test with all polytomous items and the other half were mixed format with 4 or more polytomously scored items. The number of item score levels ranged from 2 to 6.

Sample sizes for the applied studies ranged from 648 to 54,000. Reference and focal groups used were English language learners versus non English language learners with a ratio of 7:1 for 5^{th} grade and 13:1 for 8^{th} grade (Gattamorta & Penfield, 2012), English versus French speakers with a ratio of 1.6:1 (Gattamorta, Penfield, & Meyers, 2012), classrooms using a particular teaching strategy versus not with ratios of 34:1 and 36:1 for students and teachers, respectively in an Academic Forum and 16:1 and 14:1 for students and teacher, respectively in an Applied Forum (Miller et al., 2010), Hispanics versus Blacks with a ratio of about 1.2:1 (Penfield, 2008).

Methods used include the IRT likelihood ratio test using the partial credit model (in WINSTEPS) for DIF effects paired with the cumulative log odds ratio (CU-LOR) for DSF effects (Gattamorta, Penfield, & Meyers, 2012). The Mantel test, Liu-Agresti statistic, and Cox's β statistics have been used in a study to detect DIF effects while the CU-LOR statistic was used for detecting DSF effects (Miller et. al, 2010). Penfield (2008) utilized the SSL test to detect global DIF while Mantel was used for net DIF and this was paired with the CU-LOR test for DSF effects. Finally, Gattormata & Penfield (2012) used the AC-LOR and CU-LOR methods to detect DSF effects. Gattormata & Penfield (2012) mentioned inconsistent behavior with AC-LOR for their study and recommended CU-LOR; although this has not been confirmed in other studies. It is noticeable that applied studies tend to use CU-LOR instead of AC-LOR. DSF effects that were found ranged from 0 to 2. Typically, for CU-LOR, effects of 0.5 and greater were flagged as

significant. Results were not as consistent with the AC-LOR test, most likely because AC-LOR does not use as much of the sample as CU-LOR when estimating effects.

Summary of Literature Review

Recent studies have explored within item invariance for polytomous items, differential step functioning (DSF), which may be helpful to investigate if differential item functioning (DIF) effects that are opposite in sign or magnitude cause item bias to go undetected. Additionally, DSF effects give more information on which parts of a multi-step/constructed response item may need improvement.

There are few simulation and applied studies investigating the utility of differential step and item functioning together. Simulation studies have examined factors such as impact, DSF/DIF effect size, and DSF/DIF pattern on power and type I error of these methods. Regarding Type I error, factors known to inflate error rates include impact, multiple significance testing of items or item steps, and multiple items with DIF/DSF effects. Purification may have limitations in practice, thus, statistical adjustments for inflated Type I error rates may provide another solution, but few studies have explored this solution for detecting DIF in polytomous items. Regarding power, large effect sizes of DIF/DSF and constant DIF patterns tend to yield the highest power of detection. However, little is known about the effect of increasing the number of item score levels and unequal sample size ratios on DSF methods or how DSF methods compare with several DIF methods when the score levels, generating model, ability distributions or DSF/DIF patterns vary.

Based on findings in the literature, the next chapter will describe a study investigating the effects of polytomous item features on the power and Type I error of nonparametric observed score invariance tests of differential step and item functioning. Nonparametric tests are of interest due to the restrictions that parametric tests may impose, particularly sample size and model fit. The number of item score levels will be the primary effect investigated under various conditions among the following contingency table based methods: DSF tests of invariance--adjacent category and cumulative log odds ratio estimators--and DIF tests of invariance--Mantel Test, Liu Agresti, Generalized Mantel-Haenszel [GMH], and Simultaneous Step Level test [SSL].

These nonparametric methods were chosen because a) research on DSF is minimal thus far ; b) there are fewer studies on the performance of Mantel-Haenszel based DIF procedures for polytomous items than for dichotomous items; and c) the relative performance of SSL or Liu-Agresti methods to other procedures has not been fully explored. Aditionally, this study will examine which statistical procedures are most effective for adjusting per comparison Type I errors for DIF detection in polytomous items: Bonferroni, Benjamini and Hochberg, or Holm's. For the DSF/DIF methods utilized in this study. The results should help analysts better understand a) if increasing the number of item score levels when constructing multi-step/constructed response problems affect DIF detection, b) how statistical adjustments for Type I error affect DIF detection, and c) the advantages/disadvantages of using DSF methods, DIF methods, or both under various conditions for more efficient statistical detection of item bias.

CHAPTER 3

METHODS

This chapter describes the simulation methods to be used to answer the research questions described. The research questions to be addressed by this study are reprinted below for convenience.

Research questions:

- In terms of Type I error and power, which non-parametric test of invariance performs better as the number of response categories increase for polytomous items among the following methods: DSF tests of invariance (adjacent category and cumulative log odds ratio estimators) or DIF tests of invariance (Mantel Test, Liu Agresti, Generalized Mantel-Haenszel, Simultaneous Step Level test)?
- Of the Bonferroni, Benjamini and Hochberg, and Holm's methods, which procedure works best for controlling Type I error in the SSL method due to multiple significance test of DIF for polytomous items?
- Do differences in the generating model, sample size ratio, ability distributions, and pattern of DSF affect the power of DSF and DIF tests as the number of score levels increase?

The research design will consist of evaluating Type I error and power of detecting item and step level invariance for polytomous items with varying number of response options under several conditions. To answer the research questions, in the first phase of the study, a total of 48 conditions will be investigated: 2 generating models (partial credit versus graded response model) x 2 conditions of group mean difference (no impact versus impact) x 3 conditions of DSF pattern (none, convergent, divergent) x 2 conditions of number of item score levels (3 versus 4 score levels) and 2 sample size ratios (1:1 versus 5:1). In the second phase, the Bonferroni, BH, and Holm's methods will be applied to the significance testing of the SSL method.

Data Generation

Polytomously scored data will be generated using two different probability response models: the graded response model (GRM; Samejima, 1997) and the partial credit model (PCM). For the graded response model, define Y as a random variable that can take on one of a possible J ordered scores $j = 0,1, ..., J$ for an item. Now, let θ represent an examinee's latent ability being measured by the item. Then, the GRM is defined as

$$\begin{aligned} P(Y_i \geq 0 \mid \theta) &= 1 \\ P(Y_i \geq j \mid \theta) &= \frac{\exp(a(\theta - b_{ij}))}{1 + \exp(a(\theta - b_{ij}))} \qquad j = 1, ..., J, \end{aligned} \tag{18}$$

where a is the discrimination parameter common to all categories and b_j is the location parameter for category $j = 1, 2, ..., J$. In this study, a will be set at a constant value of 1 for all items. Under the PCM, the probability of scoring x on item i with $j + 1$ (from 0 to J) categories for an examinee with latent ability θ is

$$\begin{aligned} P(Y_i = 0 \mid \theta) &= 1 \\ P(Y_i = x \mid \theta) &= \frac{\exp \sum_{j=0}^{x} (\theta - b_{ij})}{\sum_{r=0}^{J} \exp \sum_{j=0}^{r} (\theta - b_{ij})}, \qquad j = 1, ..., J, \end{aligned} \tag{19}$$

where b_{ij} represents the step difficulty of item i for the category score j. The category response functions are found by taking the difference between cumulative response functions. For the GRM we have:

$$\begin{aligned} P(Y_i \geq j \mid \theta) &= P(Y_i \geq y_j \mid \theta) - P(Y \geq y_{j+1} \mid \theta) \\ P(Y = J \mid \theta) &= P(Y \geq J \mid \theta) \end{aligned} \tag{20}$$

For the PCM, we have:

$$P(Y = j \mid \theta) = P(Y_i = j \mid \theta) - P(Y = j+1 \mid \theta) \tag{21}$$

To simulate item responses, first, the probability of responding in score category j or higher on an item *i* for each examinee will be computed using the item parameters and generated ability parameters. Then, a uniform random variable, *x*, from the distribution Uniform(0,1) will be generated for each examinee and for each item. The generated random variable will be compared with the calculated model probabilities. If the generated random variable is less than the calculated probability at the score category *j* but greater than the calculated probability at the score category *j*– 1, then the response will be coded as *j*– 1.

Replications

To ensure stable results, 1000 replications will be used in this study. Previous DIF studies have used 100, 200, 300, 400, 500, and 1000+ replications (Atar, 2007; Gattamorta, 2012; Kim, 2010; Penfield, 2008; Wang & Su, 2004; Woods, 2011). Due to increase in computational power, more recent studies have used 1000 replications. Therefore, 1000 replications seem reasonable for this study.

Study Conditions

Sample Size And Sample Size Ratio

The effects of sample size has been well researched in the DIF literature. A small sample size could cause poor parameter estimation, resulting in true DIF not being detected. On the contrary, a large sample size may lead to oversensitivity of DIF detection, resulting in flagging items with very little or no DIF. Although, detection methods become very powerful for detecting items with small DIF for large sample sizes, it may be helpful to use effect sizes to determine if DIF is practically significant enough that items should be improved or removed. In DIF simulation studies, sample size usually ranges from 100 – 5000. For polytomous items, the smallest sample size

simulated has been 440 examinees with 400 in the reference group and 40 in the focal group (Woods, 2011). Thus, for this study, examinee sample size will be held constant at 1,200. Two sample size ratios will be used: 1:1 (600/600), which may represent males versus females, and 5:1 (1000/200), which may represent a majority group versus a minority group.

Ability Distributions

To simulate focal and reference group examinees, ability values will be simulated. In the DIF literature, differences between mean distribution values were typically set to be 0, 0.5, 0.75, 1 and 1.5; typically, mean differences between minority and majority groups can be higher than 0.5 (Penfield, 2008; Raju, Fortmann-Johnson, Kim, Morris, Nering, Oshima, 2009; Su & Wang, 2005; Woods, 2009). Values for the simulated reference group and focal group will be derived in one of two ways: (a) both groups come from a standard normal distribution with mean of zero and variance of one or (b) the reference group comes from a standard normal distribution and the focal group comes from a normal distribution with mean -0.75 and variance 1. All items will be simulated with a threshold mean of 0.0. Condition (a) represents equal ability between the reference and focal groups. Condition (b) represents *impact,* (Dorans & Holland, 1993) where the mean ability of the focal group differs from the mean ability of the reference group.

Score Levels

Studies specifically involving DIF in polytomous items have utilized items with about 3 to 5 categories, on average. This study will look at the effect of increasing the number of score levels on performance of DSF and DIF procedures. Items will be generated with 3 and 4 score levels. The increase in number of score levels will be of interest due to the number of significance tests needed for each additional score level and due to the reduced power to detect differences between categories with potentially less responses.

DSF conditions

Simulated examinees will take an exam consisting of 18 core items and 2 studied items. Recent studies for DIF and DSF in polytomous items considered test lengths of 8, 10, 20, 30, 40, or 50 with the number of categories ranging from 3 to 9 (although most studies use 3-5 categories as previously mentioned) (Gattamorta & Penfield 2012;

Gattamorta, Penfield, & Meyers, 2012; Wood, 2011). DIF was found to range from 1-60% in most application studies; however, simulation studies for polytomous items typically designate between 10-20% of items as containing DIF, with very few designating as high as 40-50% (Gattamorta & Penfield 2012; Gattamorta, Penfield, & Meyers, 2012; Woods, 2011; Penfield 2008; Wang & Su, 2004). In particular, applied examples of DSF studies typically find about 5 – 50% of item steps exhibiting a moderate (an effect greater than or equal to .43) to large bias effect (an effect greater than or equal to .64); for items with 3-5 categories, generally at least 1 step exhibited a large DSF effect (Gattormata & Penfield, 2012). Thus, it seems reasonable to generate a 20 item test of which two items will exhibit DIF; for each of the two items, the first step will exhibit a medium DSF effect and the second step will exhibit a large DSF effect.

The first two items of the 20 item test will be the studied items. To create DSF for the focal group, DSF effects, ω_{ij}, will be added to the b_{ij} parameters of the first two items for the reference group. The item parameters and DSF effects for the DSF conditions are shown in Tables 3 - 6. There will be a no DSF condition in which all DSF effects are zero. For the DSF effect, as the number of score levels increase from 3 to 4, the first two item steps of the studied items will exhibit either convergent DSF favoring the reference group, or divergent DSF where the reference group is favored in one step and the focal group in the other. Under the convergent condition, the DSF effects for item 1 will be $\{\omega_{11} = 0.6,\ \omega_{12} = 1.0\}$; for item 2 in the convergent condition, the DSF effects will be $\{\omega_{21} = 0.6,\ \omega_{22} = 0.8\}$. Under the divergent condition, the DSF effects for item 1 will be $\{\omega_{11} = -0.6,\ \omega_{12} = 1.0\}$; for item 2 in the divergent condition, the DSF effects will be $\{\omega_{21} = -0.6,\ \omega_{22} = 0.8\}$. The pattern of DSF for the group of studied items in each condition has been used in previous simulation studies and resembles what would be seen in application studies (e.g. Penfield & Gattamorta, 2009).

Table 3. Parameters for items with three score levels, convergent DSF condition

Item #	b_{i1}*	b_{i2}	ω_{i1}**	ω_{i2}
1	**-1.00**	**1.00**	**0.60**	**1.00**
2	**-1.00**	**1.00**	**0.60**	**0.80**
3	-1.00	0.50		
4	-1.50	2.00		
5	-1.00	1.00		
6	-2.00	1.00		
7	-0.50	1.50		
8	-1.00	0.50		
9	-1.50	2.00		
10	-1.00	1.00		
11	-2.00	1.00		
12	-0.50	1.50		
13	-1.00	0.50		
14	-1.50	2.00		
15	-1.00	1.00		
16	-2.00	0.50		
17	-0.50	2.00		
18	-1.50	1.00		
19	-1.00	1.50		
20	-2.00	0.50		

* item step

** DSF effect value for each step (positive values favor the reference group, effect values of zero are only shown for studied items.

Table 4. Parameters for items with four score levels, convergent DSF condition

Item #	b_{i1}	b_{i2}	b_{i3}	ω_{i1}	ω_{i2}	ω_{i3}
1	**-1.00**	**0.00**	**1.00**	**0.60**	**1.00**	**0.00**
2	**-1.00**	**0.00**	**1.00**	**0.60**	**0.80**	**0.00**
3	-1.00	0.00	0.50			
4	-1.50	-0.50	2.00			
5	-1.00	0.50	1.00			
6	-2.00	-0.75	1.00			
7	-0.50	0.75	1.50			
8	-1.00	0.00	0.50			
9	-1.50	-0.50	2.00			
10	-1.00	0.50	1.00			
11	-2.00	-0.75	1.00			
12	-0.50	0.75	1.50			
13	-1.00	0.00	0.50			
14	-1.50	-0.50	2.00			
15	-1.00	0.50	1.00			
16	-2.00	0.50	1.00			
17	-0.50	0.50	1.00			
18	-1.50	0.50	1.00			
19	-1.00	0.50	1.00			
20	-2.00	0.50	1.00			

Table 5. Parameters for items with three score levels, divergent DSF condition

Item #	b_{i1}	b_{i2}	ω_{i1}	ω_{i2}
1	**-1.00**	**1.00**	**-0.60**	**1.00**
2	**-1.00**	**1.00**	**-0.60**	**0.80**
3	-1.00	0.50		
4	-1.50	2.00		
5	-1.00	1.00		
6	-2.00	1.00		
7	-0.50	1.50		
8	-1.00	0.50		
9	-1.50	2.00		
10	-1.00	1.00		
11	-2.00	1.00		
12	-0.50	1.50		
13	-1.00	0.50		
14	-1.50	2.00		
15	-1.00	1.00		
16	-2.00	0.50		
17	-0.50	2.00		
18	-1.50	1.00		
19	-1.00	1.50		
20	-2.00	0.50		

Table 6. Parameters for items with four score levels, divergent DSF condition

Item #	b_{i1}	b_{i2}	b_{i3}	ω_{i1}	ω_{i2}	ω_{i3}
1	**-1.00**	**0.00**	**1.00**	**-0.60**	**1.00**	**0.00**
2	**-1.00**	**0.00**	**1.00**	**-0.60**	**0.80**	**0.00**
3	-1.00	0.00	0.50			
4	-1.50	-0.50	2.00			
5	-1.00	0.50	1.00			
6	-2.00	-0.75	1.00			
7	-0.50	0.75	1.50			
8	-1.00	0.00	0.50			
9	-1.50	-0.50	2.00			
10	-1.00	0.50	1.00			
11	-2.00	-0.75	1.00			
12	-0.50	0.75	1.50			
13	-1.00	0.00	0.50			
14	-1.50	-0.50	2.00			
15	-1.00	0.50	1.00			
16	-2.00	0.50	1.00			
17	-0.50	0.50	1.00			
18	-1.50	0.50	1.00			
19	-1.00	0.50	1.00			
20	-2.00	0.50	1.00			

Analysis

As mentioned, 1000 replications for each condition will be generated. The adjacent category log odds ratio (AC-LOR) estimator and the cumulative log odds ratio (CU-LOR) estimator will be used to detect DSF for polytomous items generated under both the PCM and GRM as an indicator of performance with model fit/misfit. Independent variables for this study are sample size ratio, impact, number of score levels, DSF pattern, and the model used to generate data. Dependent variables are Type I error

and power.

To address research question 1, for Type I error, the occurrences of falsely identified non-DIF items and non-DSF steps (false positives) will be counted per replication. The proportion of false positives will be calculated for each test per replication and will be averaged across replications. Ideally, if a significance level of $\alpha = .05$ is chosen, the Type I error rate should be approximately .05. For 1000 replications, the margin of error for this study will be calculated at a 95% confidence interval using the formula $.05 \pm 1.96\sigma/\sqrt{n}$, where $\sigma = \sqrt{p(1-p)}$ (standard deviation of a proportion), $n = 1200$ and $p = .05$; Type I error rates within the range of (0.04, 0.06) will be acceptable for the power analysis.

For methods where Type I error rates are within the simulated margin of error, power will be calculated. For each replication, power will be determined by the proportion of true positives (DSF items correctly identified) and these proportions will be averaged across replications. To determine the margin of error (using a 95% confidence interval), it is assumed that true power is 0.70; by estimating power at this level, the estimate of the margin of error will be maximized (Woods, 2011). Similarly to Type I error, if 1000 replications are used to estimate a power level of $\beta = .70$, the margin of error for a 95% confidence interval will be calculated by $.70 \pm 1.96\sigma/\sqrt{n}$, where $\sigma = \sqrt{p(1-p)}$, $n = 1200$ and $p = 0.70$. Thus, for power higher than .70, the margin of error will be less than +/- 0.03. The methods will be compared to determine which have the highest power.

To address research question 2, three methods of statistical adjustments will be compared to account for the possibility of inflated Type I error rates that may occur with multiple testing for DIF at the score level. The adjustments will be: Dunn-Bonferroni method, Benjamini-Hochberg (BH) method, and Holm's method. These adjustments have been utilized in previous DIF research, but not often. As mentioned, Dunn-Bonferroni is the most commonly used, although Benjamini-Hochberg has been recommended by What Works Clearinghouse (2014).

For research question 3, an analysis of variance (ANOVA) will be used to identify if there are any statistically significant two-way interactions between the number of score levels and either the generating model, sample size ratio, impact, or pattern of DSF effect.

Significance will be based on p-values less than .01 and effect size, η^2, greater than 0.2. For ANOVA, an effect greater than .14 is considered large (Miles & Shelvin, 2001; Cohen, 1988); thus considering effects greater than 0.2 is appropriate. The model will contain all main effects listed and all possible two-way interactions. Higher order interactions will be aggregated into the error term.

CHAPTER 4

RESULTS

Research Question 1: Type I Error Rates and Tests of Invariance

The purpose of the first research question was to investigate which non-parametric test of invariance performs better as the number of response categories increase for polytomous items for the simulated test at the alpha = 0.05 significance level. If the Type I error rate obtained by the simulation was not within the error limits for alpha, the statistical power calculations based on alpha = 0.05 may not be accurate. Again, the following methods were investigated: DSF tests of invariance (adjacent category and cumulative log odds ratio estimators) and DIF tests of invariance (Mantel Test, Liu Agresti, Generalized Mantel-Haenszel, Simultaneous Step Level test).

Type I error rates were calculated by assuming that the parameters of the studied items were identical for both the focal and reference groups. Within each condition 1,000 trials were run and six outcome measures were computed. The first two measures were the proportion of trials for which the null hypothesis of no DSF was rejected was computed for each step of each item (i.e. $z(\hat{\boldsymbol{\lambda}}_1)$, $z(\hat{\boldsymbol{\lambda}}_2)$, ..., $z(\hat{\boldsymbol{\lambda}}_j)$) based on adjacent category (AC-LOR) and cumulative category log odds (CU-LOR) ratios. The next four measures were the proportion of trials for which the null hypothesis of no DIF was rejected using the Mantel test , Liu Agresti, Generalized Mantel-Haenszel (GMH), and the Simultaneous Step Level test (SSL). DSF detection methods can be directly compared to each other and DIF methods can be directly compared to each other. Because DSF detection methods test the null hypothesis of no DSF in a particular step while the DIF methods tests the null hypothesis of no DIF for an item, they cannot be directly compared. However, since the SSL test uses the results of the DSF (CU-LOR) test to test for item DIF, it is the best way to compare the DSF framework of detecting item DIF against traditional methods. As a reminder, SSL is a multiple comparison test, that is, it rejects the null hypothesis of no DIF if any one of the step-level tests yields a significant result. Therefore, it must be adjusted to compare its results to the other DIF methods. The Dunn-Bonferroni adjustment will be used in this section based on (Penfield,

2007); however, results from the second research question will cover other statistical adjustments that could be used. Penfield (2007) adjusted the step level (DSF) Type I error rates to correct SSL Type I error rates; however, in this study the SSL Type I error rates were adjusted directly. To understand the rejection rates, if the studied item step does not have DSF, the rejection rate gives Type I error for that step. On the contrary, if that studied item step does have DSF, the rejection rate gives power for that step.

For DIF detection methods, rejection rates will also be presented for each studied item; additionally, Type I error will be averaged across DIF-free items to yield average Type I error. Type I error rates that were larger than .06 are highlighted in boldface and italic (these error limits were shown in Chapter 3). Type I error rates less than .04 are shown in bold. These conservative and liberal Type I error rates represent values that are beyond what would be expected based on random sampling. Power will be averaged across DIF items to yield average power. In the simulation, the statistical power rates for two specific DSF patterns were considered. First convergent DSF was defined by adding the constant 0.6 to the reference group's b_1 parameter for both studied items; the constant 1.0 was added to the reference groups b_2 parameter for studied item 1 and 0.8 was added to the reference group's b_2 parameter for studied item 2. Second, divergent DSF was defined by adding the constant 0.6 to the focal group's b_1 parameter for both studied items; the constant 1.0 was added to the reference groups b_2 parameter for studied item 1 and 0.8 was added to the reference group's b_2 parameter for studied item 2.

Table 7 presents rejection rates for adjacent category (AC-LOR) and cumulative log odds ratio (CU-LOR) DSF methods when no DSF was present. Seven rates fell outside of the expected interval for Type I error (0.04, 0.06) for AC-LOR; of the seven rates, six were inflated. Five rates were outside of expectation for CU-LOR, of the five rates, three were inflated. For CU-LOR, unacceptable Type I error rates occurred when the PCM model was used. There is no evidence that having three versus four item score levels made a difference in DSF detection.

Table 8 displays rejection rates for the AC-LOR and CU-LOR statistics with convergent DSF. Values in bold and italic represent conditions that had unacceptable Type I error rates for which the corresponding power rates should not be interpreted. Table 8 shows that the pattern for AC-LOR DSF detection under the PCM was low power

for the first step (DSF effect: item 1, step 1 = item 2, step 1 = 0.6), higher power for second step (DSF effect: item 1, step 2 = 1.0; item 2, step 2 = 0.8) and an appropriate Type I error rate for items with a third step which did not have DSF. Under the GRM model, the pattern for the AC-LOR statistic was similar except for inflated Type I error rates on the third item step. Table 8 shows that the pattern for the CU-LOR DSF detection under the PCM with equal sample size ratio was adequate power for the first two steps and inflated Type I error for the third step. When the sample size ratio was unequal, the CU-LOR statistic had low power in detecting the first step. Table 8 also shows that the pattern for CU-LOR DSF detection under the GRM was low power for the first step (less than 0.7) high power for the second step (greater than 0.7) and controlled Type I error for the third step. If an item had DSF in all steps, CU-LOR generally had higher power. If an item did not have DSF in all steps, AC-LOR had inflated Type I error rates for non DSF steps when GRM model is used.

Table 7. Step level rejection rates of studied items, no impact, no DSF

Model	Sample Size	Item*	AC-LOR 3 levels	AC-LOR 4 levels	CU-LOR 3 levels	CU-LOR 4 levels
PCM	600/600	Item 1				
		Step 1	***0.061***	0.047	0.058	**0.037**
		Step 2	0.060	0.047	***0.063***	0.052
		Step 3		0.047		0.051
		Item 2				
		Step 1	0.049	***0.061***	0.045	***0.063***
		Step 2	0.042	0.042	0.045	0.043
		Step 3		0.052		0.052
	1000/200	Item 1				
		Step 1	0.045	0.043	0.042	0.044
		Step 2	**0.030**	0.058	**0.031**	0.049
		Step 3		0.042		0.045
		Item 2				
		Step 1	0.058	***0.061***	***0.064***	0.056
		Step 2	0.036	0.045	0.038	0.043
		Step 3		0.049		0.052
GRM	600/600	Item 1				
		Step 1	0.051	0.051	0.048	0.043
		Step 2	0.046	0.044	0.049	0.052
		Step 3		0.055		0.052
		Item 2				
		Step 1	0.050	0.055	0.053	0.049
		Step 2	0.042	0.036	0.044	0.045
		Step 3		0.051		0.045
	1000/200	Item 1				
		Step 1	0.048	***0.062***	0.050	0.051
		Step 2	0.042	0.046	0.044	0.047
		Step 3		0.050		0.054
		Item 2				
		Step 1	0.058	***0.061***	0.052	0.050
		Step 2	***0.070***	0.052	0.053	0.051
		Step 3		0.046		0.043

*DSF effects are all zero. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 8. Step level rejection rates of studied items with no impact and convergent DSF

Model	Sample Size	Item*	AC-LOR 3 levels	AC-LOR 4 levels	CU-LOR 3 levels	CU-LOR 4 levels
PCM	600/600	Item 1				
		Step 1 (0.6)	***0.381***	0.264	0.717	**0.811**
		Step 2 (1.0)	0.952	0.925	***0.982***	0.991
		Step 3 (0.0)		0.049		***0.236***
		Item 2				
		Step 1 (0.6)	0.357	***0.276***	0.639	***0.715***
		Step 2 (0.8)	0.877	0.835	0.948	0.968
		Step 3 (0.0)		0.048		***0.159***
	1000/200	Item 1				
		Step 1 (0.6)	0.247	***0.140***	0.487	0.544
		Step 2 (1.0)	**0.829**	0.736	**0.902**	0.910
		Step 3 (0.0)		0.049		***0.166***
		Item 2				
		Step 1 (0.6)	0.202	0.161	***0.393***	0.460
		Step 2 (0.8)	0.642	0.552	0.752	0.796
		Step 3 (0.0)		0.051		***0.117***
GRM	600/600	Item 1				
		Step 1 (0.6)	0.057	0.198	0.449	0.502
		Step 2 (1.0)	0.966	0.999	0.990	0.996
		Step 3 (0.0)		***0.805***		0.052
		Item 2				
		Step 1 (0.6)	0.104	0.074	0.445	0.467
		Step 2 (0.8)	0.848	0.972	0.924	0.956
		Step 3 (0.0)		***0.688***		0.060
	1000/200	Item 1				
		Step 1 (0.6)	0.057	***0.111***	0.250	0.289
		Step 2 (1.0)	0.810	0.959	0.877	0.903
		Step 3 (0.0)		***0.546***		0.053
		Item 2				
		Step 1 (0.6)	0.082	***0.063***	0.259	0.282
		Step 2 (0.8)	***0.633***	0.821	0.743	0.776
		Step 3 (0.0)		***0.446***		0.065

*DSF effects in parentheses. Italic rates- Type I error < .04; Bold and italic rates- Type I error > .06

CU-LOR will have inflated Type I error rates for the non DSF steps, when the PCM model is used. In general, when Type I error rates were inflated, AC-LOR had higher inflation than CU-LOR.

Table 9 reveals that for the divergent DSF pattern, the CU-LOR statistic had more controlled Type I error rates when the GRM model was used and the AC-LOR statistic had more controlled Type I error rates when the PCM model was used. Generally, the CU-LOR statistic had low power on the first step (DSF effect: item 1, step 1 = item 2, step 1 = 0.6) and high power for the second step (DSF effect: item 1, step 2 = 1.0; item 2, step 2 = 0.8). AC-LOR statistic had a similar pattern for power except under the GRM model, power was very high for steps containing DSF. Although not shown here, the effect of impact on rejection rates of the differential step functioning procedures was consistent and will now be discussed briefly. In general, the pattern of rejection rates were unchanged (i.e. lower rate for the first step, higher rate for the second step); there were a few changes in the magnitude of the rejection rates. The convergent condition yielded slightly higher power rates for items with 4 score levels and 1: 1 sample size ratio as well as inflated rates for the non-DSF step. The divergent condition yielded slightly lower power rates for items with 3 score levels when the GRM model was used and the sample size ratio was 5:1.

Regarding DIF detection methods, average Type I error and power rates will be presented first, followed by rejection rates for each studied item. Table 10 displays average Type I error rates for the four DIF detection methods. When no DSF were present, the Type I error was within the error limits of .05 +/- .01 for all methods.

Table 11 presents results with the convergent DSF pattern. Out of the DIF detection methods, the SSL statistic had the best control over Type I error with only four rates falling outside the expected interval of (0.04. 0.06). Problems mostly occurred with three item score levels with equal sample size. GMH statistic had the second best control over Type I error. The generalized Mantel-Haenzel GMH method had six Type I error rates, ranging from .062 to .088, that fell out of the predicted interval. These rates occurred when items had 3 score levels and the sample size ratio was equal; when items had 4 score levels, Type I error rates were inflated for GMH if sample size were equal and the PCM model was used. The Mantel method had ten Type I error rates that fell

Table 9. Step level rejection rates of studied items with no impact and divergent DSF

	Sample		AC-LOR		CU-LOR	
Model	Size	Item	3 levels	4 levels	3 levels	4 levels
PCM	600/600	Item 1				
		Step 1 (0.6)	***0.475***	0.347	***0.183***	***0.052***
		Step 2 (1.0)	0.966	0.937	0.935	0.922
		Step 3 (0.0)		0.057		***0.191***
		Item 2				
		Step 1 (0.6)	0.489	***0.357***	0.253	***0.078***
		Step 2 (0.8)	0.875	0.829	0.792	0.757
		Step 3 (0.0)		0.058		***0.120***
	1000/200	Item 1				
		Step 1 (0.6)	0.300	***0.220***	0.126	0.049
		Step 2 (1.0)	**0.816**	0.746	**0.761**	0.737
		Step 3 (0.0)		0.049		***0.116***
		Item 2				
		Step 1 (0.6)	0.304	0.180	***0.158***	0.061
		Step 2 (0.8)	0.641	0.570	0.554	0.509
		Step 3 (0.0)		0.044		***0.089***
GRM	600/600	Item 1				
		Step 1 (0.6)	0.995	1.000	0.632	0.652
		Step 2 (1.0)	1.000	1.000	0.998	0.997
		Step 3 (0.0)		***0.795***		0.050
		Item 2				
		Step 1 (0.6)	0.985	1.000	0.651	0.634
		Step 2 (0.8)	0.998	1.000	0.956	0.975
		Step 3 (0.0)		***0.795***		***0.062***
	1000/200	Item 1				
		Step 1 (0.6)	0.938	***1.000***	0.455	0.430
		Step 2 (1.0)	0.992	0.975	0.903	0.935
		Step 3 (0.0)		***0.559***		0.050
		Item 2				
		Step 1 (0.6)	0.878	***1.000***	0.422	0.431
		Step 2 (0.8)	***0.968***	0.999	0.794	0.811
		Step 3 (0.0)		***0.448***		0.045

*DSF effects in parentheses. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 10. Type 1 Error Rates of DSF and DIF Detection Methods for Items with No DSF

Sample Size	Model	Impact	Mantel	GMH	Liu-Agresti	SSL*
3 Item Score Levels						
600/600	pcm	no	.049	.049	.049	.046
		yes	.053	.052	.051	.051
	grm	no	.049	.048	.048	.046
		yes	.051	.052	.051	.049
1000/200	pcm	no	.048	.049	.048	.045
		yes	.049	.051	.049	.046
	grm	no	.052	.051	.052	.048
		yes	.052	.053	.051	.047
4 Item Score Levels						
600/600	pcm	no	.046	.051	.046	.046
		yes	.052	.050	.052	.049
	grm	no	.049	.051	.049	.045
		yes	.051	.051	.051	.043
1000/200	pcm	no	.046	.049	.046	.042
		yes	.051	.049	.051	.044
	grm	no	.050	.051	.049	.042
		yes	.048	.048	.048	.040

*Dunn-Bonferroni adjusted. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 11. Type 1 Error Rates of DSF and DIF Detection Methods for Items with Convergent DSF

Sample Size	Model	Impact	Mantel	GMH	Liu-Agresti	SSL*
3 Item Score Levels						
600/600	pcm	no	***.076***	***.069***	***.075***	***.069***
		yes	***.074***	***.066***	***.073***	***.063***
	grm	no	***.072***	***.067***	***.072***	***.065***
		yes	***.068***	***.062***	***.068***	.060
1000/200	pcm	no	***.063***	.058	***.062***	.055
		yes	***.063***	.058	***.062***	.056
	grm	no	.060	.057	.060	.055
		yes	.058	.056	.058	.052
4 Item Score Levels						
600/600	pcm	no	***.075***	***.062***	***.075***	.060
		yes	***.123***	***.088***	***.121***	***.091***
	grm	no	.057	.053	.057	.048
		yes	.058	.056	.058	.048
1000/200	pcm	no	***.061***	.055	***.061***	.052
		yes	***.062***	.053	***.062***	.052
	grm	no	.056	.052	.056	.046
		yes	.055	.053	.055	.046

*Dunn-Bonferroni adjusted. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

outside the predicted interval. This occurred when items had 3 score levels, except when the sample size ratio was unequal and the GRM was used; with 4 item score levels, it always occurred when the PCM was used. The Liu Agresti method yielded similar results to the Mantel method. Type I error rates with divergent DSF shown in Table 12 had a similar pattern as those shown in Table 10 with No DSF. Type I error rates were controlled.

Based on results from calculating Type I error rates, the statistical power of the Mantel test, the Liu-Agresti statistic, and the simultaneous step level (SSL) test can be evaluated. Given Type I error rates within the expected range of (0.04, 0.06), the percentage of time that the studied item score levels were correctly flagged for DSF or the studied items was correctly flagged for DIF is the estimated power rate. Generally, power rates were lower when the sample size ratio was unequal. Table 13 reveals that when comparing acceptable power rates among all methods under the convergent DSF condition (those for which Type I error rates were within the predicted interval of 0.5 +/- 0.1), Mantel had the highest power when items had 3 score levels and GMH had the highest power when items had 4 score levels. The SSL test had power rates that were similar to the GMH statistic. The Mantel test and the Liu-Agresti statistic power rates ranged from .447 to 1.00 and were either the same or usually higher than the GMH statistic; the only exceptions were when items had 4 score levels and the GRM was used. However, several rates for the Mantel test and Liu-Agresti cannot be fully interpreted due to their corresponding high

Type I error rates; this was particularly true for items with 3 score levels.

Table 14 shows that under the divergent DSF pattern, the GMH statistic and SSL test had the higher power rates. While the Mantel test and Liu-Agresti statistic had only three power rates above .70, the GMH statistic had twelve power rates above .70.

Table 12. Type 1 Error Rates of DSF and DIF Detection Methods for Items with Divergent DSF

Sample Size	Model	Impact	Mantel	GMH	Liu-Agresti	SSL*
3 Item Score Levels						
600/600	pcm	no	.049	.052	.049	.050
		yes	.047	.050	.047	.048
	grm	no	.053	.052	.053	.050
		yes	.051	.050	.051	.049
1000/200	pcm	no	.052	.052	.052	.049
		yes	.049	.050	.049	.047
	grm	no	.052	.053	.051	.049
		yes	.049	.052	.048	.046
4 Item Score Levels						
600/600	pcm	no	.054	.055	.055	.051
		yes	.056	.052	.055	.051
	grm	no	.048	.049	.048	.042
		yes	.052	.054	.052	.046
1000/200	pcm	no	.049	.051	.050	.046
		yes	.051	.048	.052	.043
	grm	yes	.052	.048	.051	.042
		no	.050	.050	.049	.042

*Dunn-Bonferroni adjusted. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 13. Power Rates of DSF and DIF Detection Methods for Items with Convergent DSF

Sample Size	Model	Impact	Mantel	GMH	Liu-Agresti	SSL*
3 Item Score Levels						
600/600	pcm	no	***.981***	***.968***	***.981***	***.960***
		yes	***.964***	***.939***	***.961***	***.929***
	grm	no	***.938***	***.937***	***.938***	***.942***
		yes	***.890***	.893	***.890***	.901
1000/200	pcm	no	***.864***	.818	***.860***	.808
		yes	.802	.752	.794	.736
	grm	no	.733	.729	.726	.742
		yes	.690	.680	.678	.692
4 Item Score Levels						
600/600	pcm	no	***.964***	***.963***	***.964***	.964
		yes	1.00	***1.00***	1.00	***1.00***
	grm	no	.693	.986	.692	.943
		yes	.685	.982	.686	.936
1000/200	pcm	no	***.808***	.766	***.806***	.778
		yes	***.852***	.780	***.852***	.788
	grm	no	.447	.864	.454	.722
		yes	.470	.884	.474	.730

*Dunn-Bonferroni adjusted. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 14. Power Rates of DIF Detection Methods for Items with Divergent DSF

Sample Size	Model	Impact	Mantel	GMH	Liu-Agresti	SSL*
3 Item Score Levels						
600/600	pcm	no	.259	.914	.258	.850
		yes	.116	.846	.114	.756
	grm	no	.192	1.00	.192	.994
		yes	.156	1.00	.158	.987
1000/200	pcm	no	.190	.691	.181	.601
		yes	.086	.607	.078	.488
	grm	no	.142	.986	.124	.908
		yes	.110	.970	.094	.856
4 Item Score Levels						
600/600	pcm	no	.447	.824	.449	.850
		yes	.330	.776	.330	.756
	grm	no	.138	1.00	.137	.994
		yes	.115	1.00	.120	.987
1000/200	pcm	no	.286	.559	.278	.488
		yes	.188	.523	.177	.413
	grm	no	.098	1.00	.090	.936
		yes	.090	1.00	.089	.925

Tables 15-17 show item level rejection rates for the DIF methods with no impact. Table entries for the power analysis when Type I errors were out of range were highlighted because power results are may not be as reliable. When no DSF was present, only the GMH statistic exhibited inflated Type I error (0.064) for a 3 score level item under the PCM model and equal sample size. Otherwise, the DIF methods had similar rejection rates in the no DSF condition. For all DIF methods, Type I error rates were deflated (less

than 0.40) under the PCM model with unequal sample size ratio and under the GRM model with equal sample size ratio (except for the Mantel statistic which was 0.41).

With the convergent DSF condition, the Mantel had five power rates that were below the interval (0.77, 0.83) ranging from 0.38 to 0.75. When items had 4 score levels, these lower rates occurred under every condition except when the PCM was used and a) sample size ratio was 1:1 b) sample size ratio was 5:1 for the first item. When items had 3 score levels, a lower power rate occurred under the GRM model with unequal sample size ratio. The Liu-Agresti statistic exhibited similar behavior as the Mantel statistic. The GMH model had three power rates fall below .80 +/- .03 ranging from 0.66 to 0.74; this was the fewest among all DIF detection methods and occurred for item 2 when the sample size ratio was unequal, except when item 2 had 4 score levels and the GRM model was used. The SSL test had similar behavior as did GMH except it more consistently had lower rates for item 2 when sample size ratio was unequal; thus it had four rates below 0.77.

For the divergent DSF pattern, both the Mantel test and Liu-Agresti statistic had poor power rates over all conditions ranging from 0.07 to 0.58. The GMH statistic had five power rates less than 0.77, ranging from 0.47 to 0.75; these rates occurred under the PCM model for item 2 when the sample size ratio was equal and for both items when the sample size ratio was unequal. The SSL test exhibited a similar pattern of behavior as the GMH statistic; the five power rates that were less than 0.77 ranged from 0.39 to 0.70.

Table 15. Item level rejection rates of studied items with no impact and no DSF

Model	Sample Size	Item*	Mantel	GMH	Liu-Agresti	SSL*
PCM	600/600	Item 1				
		3 levels	0.055	***0.064***	0.056	0.058
		4 levels	0.053	0.048	0.052	0.044
		Item 2				
		3 levels	0.049	0.045	0.049	0.040
		4 levels	0.043	0.044	0.043	0.045
	1000/200	Item 1				
		3 levels	**0.039**	0.026	0.038	0.030
		4 levels	0.052	0.047	0.054	0.051
		Item 2				
		3 levels	0.054	0.052	0.053	0.047
		4 levels	0.053	0.055	0.056	0.047
GRM	600/600	Item 1				
		3 levels	0.041	0.038	0.039	0.039
		4 levels	0.051	0.049	0.051	0.045
		Item 2				
		3 levels	0.040	0.039	0.042	0.046
		4 levels	0.046	0.052	0.046	0.051
	1000/200	Item 1				
		3 levels	0.045	0.051	0.046	0.043
		4 levels	0.056	0.055	0.051	0.042
		Item 2				
		3 levels	0.054	***0.063***	0.055	0.055
		4 levels	0.054	0.047	0.053	0.044

*All DSF effects are zero. Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 16. Item level rejection rates of studied items with no impact and convergent DSF

Model	Sample Size	Item	Mantel	GMH	Liu-Agresti	SSL*
PCM	600/600	Item 1				
		3 levels	0.992	0.988	0.992	0.982
		4 levels	0.981	0.982	0.981	0.983
		Item 2				
		3 levels	0.970	0.948	0.970	0.938
		4 levels	0.947	0.944	0.947	0.946
	1000/200	Item 1				
		3 levels	0.919	0.895	0.914	0.886
		4 levels	0.867	0.839	0.863	0.854
		Item 2				
		3 levels	0.810	0.740	0.806	0.730
		4 levels	0.750	0.693	0.750	0.702
GRM	600/600	Item 1				
		3 levels	0.962	0.970	0.962	0.978
		4 levels	0.775	0.998	0.772	0.991
		Item 2				
		3 levels	0.913	0.904	0.915	0.907
		4 levels	0.611	0.975	0.612	0.895
	1000/200	Item 1				
		3 levels	0.788	0.798	0.782	0.817
		4 levels	0.507	0.929	0.513	0.812
		Item 2				
		3 levels	0.678	0.66	0.671	0.668
		4 levels	0.387	0.798	0.395	0.632

*DSF effects (favors reference group): a) Item 1 & 2, Step 1 = 0.6, b) Item 1, Step 2 = 1.0, c) Item 2, Step 2 = 0.8

Table 17. Item level rejection rates of studied items with no impact and divergent DSF

Sample Size	Item	Mantel	GMH	Liu-Agresti	SSL
600/600	Item 1				
	3 levels	0.358	0.953	0.357	0.923
	4 levels	0.585	0.897	0.592	0.853
	Item 2				
	3 levels	0.160	0.875	0.159	0.778
	4 levels	0.309	0.752	0.306	0.643
1000/200	Item 1				
	3 levels	0.259	0.769	0.249	0.701
	4 levels	0.359	0.644	0.347	0.580
	Item 2				
	3 levels	0.120	0.613	0.113	0.501
	4 levels	0.212	0.474	0.208	0.395
600/600	Item 1				
	3 levels	0.265	1.000	0.267	0.999
	4 levels	0.165	1.000	0.163	1.000
	Item 2				
	3 levels	0.118	1.000	0.117	0.989
	4 levels	0.111	1.000	0.111	0.996
1000/200	Item 1				
	3 levels	0.179	0.995	0.157	0.951
	4 levels	0.108	1.000	0.108	0.986
	Item 2				
	3 levels	0.106	0.976	0.091	0.865
	4 levels	0.071	1.000	0.070	0.887

*DSF effects: a) Item 1 & 2, Step 1 = -0.6, b) Item 1, Step 2 = 1.0, c) Item 2, Step 2 = 0.8

Research Question 2: Type I Error Adjustments for Multiple Significance Testing

Research question 2 asked which procedure of the Bonferroni, Benjamini and Hochberg, and Holm's methods worked best for controlling Type I error due to multiple significance tests of DIF for polytomous items. Multiple significance tests of DIF occurred when using the SSL test. Figure 1 in the appendix displays unadjusted and adjusted *p*-values for selected conditions of the cumulative category logs odds ratio (CU-LOR) statistic. Only this statistic is shown because the results from the CU-LOR statistic for DSF at each item step is used to calculate the SSL test for DIF of the item. The pattern in these figures were consistent in that 1) *p*-values were higher for all items and item steps when no DSF was present; 2) *p*-values were lowest in the conditions for which studied items contained DSF with slight variations based on DSF pattern; 3) unadjusted *p*-values were always lowest in rank, followed by Benjamini-Hochberg, Holm, and finally Dunn-Bonferroni.

Table 18 shows that all adjusted Type I error rates for the SSL method were well controlled when no DSF was present. When convergent DSF was present under the Dunn-Bonferroni adjustment and Holm adjustments, the SSL method had better control over Type I error when compared to the other DIF detection methods. The SSL method had four Type I error rates fall out of the predicted interval .05 +/- .01 (Table 19). For items having 3 score levels, the inflated rates occurred when sample size ratios were 1:1 except when the graded response model (GRM) was used and impact was present. For items having 4 score levels, the SSL method had inflated rates when sample size ratio was 1:1, and the PCM was used while impact was present. With the Benjamini-Hochberg (BH) adjustment, the SSL method had four Type I error rates fall out of the predicted interval (.040, .060) when convergent DSF was present. When items had 3 score levels, inflated rates occurred when sample size ratios were 1:1. When the items had 4 score levels, the SSL method had inflated rates when sample size ratios were 1:1 and the PCM was used. Table 20 shows that all adjusted Type I error rates for the SSL method were well controlled for divergent DSF, similar to when no DSF was present. In regards to power, rates were adequate even after adjustments were implemented except in the divergent case with a 5:1 sample size ratio and the PCM; BH adjusted power tending to have higher rates, in general. When compared to the other DIF detection methods,

Mantel, GMH and Liu-Agresti, power rates were quite similar for items with 3 score levels. However, for items with 4 score levels, SSL adjusted power rates tended to be higher than Mantel and GMH power rates for the convergent DSF pattern. All power rates for the SSL test were above .80 +/- .03 in the convergent DSF condition.

Table 18. Adjusted Type 1 Error and Power Rates for the Simultaneous Step Level (SSL) DIF test with No DSF*

Sample			Type I Error				Power		
Size	Model	Impact	Unadjust	Dunn-B	BH	Holm	Dunn-B	BH	Holm
3 Item Score Levels									
600/600	pcm	no	***.096***	.046	.047	.046	NA	NA	NA
		yes	***.098***	.051	.051	.051			
	grm	no	***.096***	.046	.047	.046			
		yes	***.097***	.049	.050	.049			
1000/200	pcm	no	***.092***	.045	.046	.045			
		yes	***.095***	.046	.047	.046			
	grm	no	***.097***	.048	.049	.048			
		yes	***.097***	.047	.048	.047			
4 Item Score Levels									
600/600	pcm	no	***.139***	.046	.047	.046			
		yes	***.137***	.049	.050	.049			
	grm	no	***.124***	.045	.048	.045			
		yes	***.126***	.043	.046	.043			
1000/200	pcm	no	***.131***	.042	.043	.042			
		yes	***.134***	.044	.045	.044			
	grm	no	***.123***	.042	.045	.042			
		yes	***.120***	.040	.043	.040			

*Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 19. Adjusted Type 1 Error and Power Rates for the Simultaneous Step Level (SSL) DIF test with Convergent DSF*

Sample			Type I Error				Power		
Size	Model	Impact	Unadjust	Dunn-B	BH	Holm	Dunn-B	BH	Holm
3 Item Score Levels									
600/600	pcm	no	***.122***	***.069***	***.070***	***.069***	***.960***	***.964***	***.960***
		yes	***.119***	***.063***	***.064***	***.063***	***.929***	***.937***	***.929***
	grm	no	***.120***	***.065***	***.066***	***.065***	***.942***	***.945***	***.942***
		yes	***.114***	.060	***.061***	.060	.901	***.906***	.901
1000/200	pcm	no	***.106***	.055	.056	.055	.808	.816	.808
		yes	***.107***	.056	.057	.056	.736	.746	.736
	grm	no	***.104***	.055	.056	.055	.742	.751	.742
		yes	***.102***	.052	.053	.052	.692	.696	.692
4 Item Score Levels									
600/600	pcm	no	***.161***	.060	***.062***	.060	.964	***.968***	.964
		yes	***.213***	***.091***	***.094***	***.091***	***1.00***	***1.00***	***1.00***
	grm	no	***.128***	.048	.052	.048	.943	.945	.943
		yes	***.135***	.048	.050	.048	.936	.939	.936
1000/200	pcm	no	***.147***	.052	.054	.052	.778	.790	.778
		yes	***.144***	.052	.053	.052	.788	.802	.788
	grm	no	***.130***	.046	.049	.046	.722	.725	.722
		yes	***.129***	.046	.049	.046	.730	.734	.730

*Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Table 20. Type 1 Error and Power Rate Adjustments for the Simultaneous Step Level (SSL) DIF Test with Divergent DSF

Sample			Type I Error				Power		
Size	Model	Impact	Unadjust	Dunn-B	BH	Holm	Dunn-B	BH	Holm
3 Item Score Levels									
600/600	pcm	no	***.096***	.050	.051	.050	.850	.855	.850
		yes	***.095***	.048	.049	.048	.756	.764	.756
	grm	no	***.097***	.050	.051	.050	.994	.995	.994
		yes	***.094***	.049	.050	.049	.987	.991	.987
1000/200	pcm	no	***.098***	.049	.050	.049	.601	.605	.601
		yes	***.094***	.047	.047	.047	.488	.496	.488
	grm	no	***.097***	.049	.050	.049	.908	.918	.908
		yes	***.097***	.046	.048	.046	.856	.869	.856
4 Item Score Levels									
600/600	pcm	no	***.140***	.051	.052	.051	.850	.753	.748
		yes	***.141***	.051	.053	.051	.756	.672	.666
	grm	no	***.123***	.042	.044	.042	.994	999	.998
		yes	***.126***	.046	.049	.046	.987	1.00	1.00
1000/200	pcm	no	***.135***	.046	.050	.046	.488	.492	.488
		yes	***.135***	.043	.047	.043	.413	.418	.413
	grm	no	***.126***	.042	.050	.042	.936	.946	.936
		yes	***.123***	.042	.048	.042	.925	.938	.925

*Italic rates- Type I error < .04; Bold and italic rates - Type I error > .06

Research Question 3: Effect of Conditions on Power (ANOVA)

The final research question for this study utilized ANOVA methodology to explore which of five factors affected the percentage of time an item was flagged for DSF/DIF: generating model used, inclusion or exclusion of impact, the ratio of the reference group to the focal group, the number of score levels for the item and the pattern of DSF. The model included all main effects and two-way interactions. Combining higher order interactions terms in the error term increases the degrees of freedom for the error term; but, mainly, there was no interest in interpreting higher order interaction based on the research questions being investigated. In the case of DSF, the dependent variable was defined as the percentage of time that item steps containing DSF was flagged for DSF. In the case of DIF, the dependent variable was defined as the percentage of time that the studied items were flagged for DIF. In both cases, studied item steps or studied items were flagged using method *M*, where method *M* was any of the six DSF or DIF detection methods presented in the first research question. The results of the six ANOVA tests are presented in Tables 21-26.

Significant interaction effects should be identified before significant main effects can be interpreted. The *p* - value indicates the probability of finding the observed result when the null hypothesis is true and tends to become smaller as sample size increases. Thus, eta squared was used as an indication of how large an effect actually is (measure of effect size) when compared to other effects. Values of eta squared greater than .14 (Cohen, 1988; Miles & Shelvin, 2001) were considered as a large effect. To be considered significant in the ANOVA results, both the *p* - value had to be less than .01 and Eta squared had to be greater than .2. No interactions met both criteria, therefore, attention will be given to the main effects. The six ANOVA tables confirm that neither the item score level effect (3 levels versus 4 levels) nor interactions involving the item score level effect had both statistical and practical significance. However, the pattern of DSF effect was both statistically ($p < .001$) and practically significant (Eta squared > 0.3) for all DSF and DIF detection methods.

Table 21. ANOVA results for statistical power rates using adjacent category log odds ratio (AC-LOR)

Source	df	SS	MS	F	p-value	Eta squared
model	1	26439	26439	2185.9	< 0.001	0.03
impact	1	12	12	1.0	0.31	0.00
ratio	1	15760	15760	1303.0	< 0.001	0.01
levels	1	7050	7050	582.9	< 0.001	0.01
pattern	2	354087	177044	14637.0	**< 0.001**	**0.34**
model x impact	1	215	215	17.8	< 0.001	0.00
model x ratio	1	1541	1541	127.4	< 0.001	0.00
model x level	1	4031	4031	333.2	< 0.001	0.00
model x pattern	2	51564	25782	2131.5	< 0.001	0.05
impact x ratio	1	273	273	22.6	< 0.001	0.00
impact x levels	1	330	330	27.3	< 0.001	0.00
impact x pattern	2	425	212	17.6	< 0.001	0.00
ratio x levels	1	12	12	1.0	0.33	0.00
ratio x pattern	2	7811	3905	322.9	< 0.001	0.01
levels x pattern	2	5838	2919	241.3	< 0.001	0.01
Residuals	47979	580334	12			

Note: ANOVA model includes all main effects and two-way interactions. Effects with both significant *p*-values and large eta squared values are bold.

Table 22. ANOVA results for statistical power rates using cumulative category log odds ratio (CU-LOR)

Source	df	SS	MS	F	p-value	Eta squared
model	1	543	543	41.2	< 0.001	0.00
impact	1	4	4	0.3	0.577	0.00
ratio	1	21158	21158	1604.7	< 0.001	0.02
levels	1	13607	13607	1032.0	< 0.001	0.01
pattern	2	333936	166968	12663.4	**< 0.001**	**0.32**
model x impact	1	328	328	24.8	< 0.001	0.00
model x ratio	1	710	710	53.8	< 0.001	0.00
model x level	1	1479	1479	112.2	< 0.001	0.00
model x pattern	2	35204	17602	1335.0	< 0.001	0.03
impact x ratio	1	305	305	23.2	< 0.001	0.00
impact x levels	1	970	970	73.6	< 0.001	0.00
impact x pattern	2	830	415	31.5	< 0.001	0.00
ratio x levels	1	71	71	5.4	0.021	0.00
ratio x pattern	2	11525	5763	437.0	< 0.001	0.01
levels x pattern	2	6318	3159	239.6	< 0.001	0.01
Residuals	47979	632608	13			

Note: ANOVA model includes all main effects and two-way interactions. Effects with both significant *p*-values and large eta squared values are bold.

Table 23. ANOVA results for statistical power rates using the Mantel Test

Source	df	SS	MS	F	p-value	Eta squared
model	1	25317	25317	854.1	< 0.001	0.01
impact	1	18	18	0.6	0.431	0.00
ratio	1	21413	21413	722.4	< 0.001	0.01
levels	1	279	279	9.4	0.002	0.00
pattern	2	762920	381460	12869.4	**< 0.001**	**0.33**
model x impact	1	175	175	5.9	0.015	0.00
model x ratio	1	1172	1172	39.5	< 0.001	0.00
model x level	1	13932	13932	470.0	< 0.001	0.01
model x pattern	2	24749	12374	417.5	< 0.001	0.01
impact x ratio	1	460	460	15.5	< 0.001	0.00
impact x levels	1	2385	2385	80.5	< 0.001	0.00
impact x pattern	2	2658	1329	44.8	< 0.001	0.00
ratio x levels	1	930	930	31.4	< 0.001	0.00
ratio x pattern	2	21770	10885	367.2	< 0.001	0.01
levels x pattern	2	5778	2889	97.5	< 0.001	0.00
Residuals	47979	1422134	30			

Note: ANOVA model includes all main effects and two-way interactions. Effects with both significant *p*-values and large eta squared values are bold.

Table 24. ANOVA results for statistical power rates using the Generalized Mantel Haenszel statistic (GMH)

Source	df	SS	MS	F	p-value	Eta squared
model	1	6250	6250	230.3	< 0.001	0.00
impact	1	124	124	4.6	0.033	0.00
ratio	1	22127	22127	815.4	< 0.001	0.01
levels	1	23	23	0.9	0.353	0.00
pattern	2	800430	400215	14749.0	**< 0.001**	**0.37**
model x impact	1	0	0	0.0	0.937	0.00
model x ratio	1	4775	4775	176.0	< 0.001	0.00
model x level	1	523	523	19.3	< 0.001	0.00
model x pattern	2	25634	12817	472.3	< 0.001	0.01
impact x ratio	1	298	298	11.0	< 0.001	0.00
impact x levels	1	675	675	24.9	< 0.001	0.00
impact x pattern	2	687	343	12.7	< 0.001	0.00
ratio x levels	1	134	134	5.0	0.026	0.00
ratio x pattern	2	13656	6828	251.6	< 0.001	0.01
levels x pattern	2	1504	752	27.7	< 0.001	0.00
Residuals	47979	1301914	27			

Note: ANOVA model includes all main effects and two-way interactions. Effects with both significant *p*-values and large eta squared values are bold.

Table 25. ANOVA results for statistical power rates using the Liu-Agresti statistic

Source	df	SS	MS	F	p-value	Eta squared
model	1	25114	25114	852.8	< 0.001	0.01
impact	1	29	29	1.0	0.321	0.00
ratio	1	22950	22950	779.3	< 0.001	0.01
levels	1	146	146	5.0	0.026	0.00
pattern	2	761944	380972	12937.1	**< 0.001**	**0.33**
model x impact	1	131	131	4.5	0.035	0.00
model x ratio	1	853	853	29.0	< 0.001	0.00
model x level	1	14018	14018	476.0	< 0.001	0.01
model x pattern	2	23816	11908	404.4	< 0.001	0.01
impact x ratio	1	361	361	12.2	< 0.001	0.00
impact x levels	1	2376	2376	80.7	< 0.001	0.00
impact x pattern	2	2523	1262	42.8	< 0.001	0.00
ratio x levels	1	758	758	25.7	< 0.001	0.00
ratio x pattern	2	21278	10639	361.3	< 0.001	0.01
levels x pattern	2	5738	2869	97.4	< 0.001	0.00
Residuals	47979	1412883	29			

Note: ANOVA model includes all main effects and two-way interactions. Effects with both significant *p*-values and large eta squared values are bold.

Table 26. ANOVA results for statistical power rates using the Simultaneous Step Level (SSL) Test

Source	df	SS	MS	F	p-value	
model	1	1355	1355	25.911	< 0.001	0.00
impact	1	2	2	0.047	0.828	0.00
ratio	1	25064	25064	479.254	< 0.001	0.01
levels	1	137482	137482	2628.867	< 0.001	0.04
pattern	2	830603	415302	7941.219	**< 0.001**	**0.23**
model x impact	1	137	137	2.621	0.105	0.00
model x ratio	1	4855	4855	92.827	< 0.001	0.00
model x level	1	8379	8379	160.224	< 0.001	0.00
model x pattern	2	27567	13784	263.565	< 0.001	0.01
impact x ratio	1	792	792	15.141	< 0.001	0.00
impact x levels	1	1870	1870	35.764	< 0.001	0.00
impact x pattern	2	1166	583	11.148	< 0.001	0.00
ratio x levels	1	200	200	3.816	0.051	0.00
ratio x pattern	2	12867	6433	123.018	< 0.001	0.00
levels x pattern	2	581	291	5.557	0.004	0.00
Residuals	47979	2509156	52			

Note: ANOVA model includes all main effects and two-way interactions. Effects with both significant *p*-values and large eta squared values are bold.

Table 27. Means and Standard Deviations for Statistical Power using Adjacent Category Log Odds Ratio (AC-LOR)

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	5.0	3.2
Divergent	16000	11.4	5.9
Convergent	16000	9.7	5.1
Item Score Levels			
3 levels	24000	9.1	5.8
4 levels	24000	8.3	5.1
Sample Size Ratio			
600/600	24000	9.2	5.7
1000/200	24000	8.1	5.2
Impact			
No Impact	24000	8.7	5.5
Impact Present	24000	8.7	5.5
Model			
PCM	24000	7.9	5.0
GRM	24000	9.4	5.9

Descriptive statistics for the ANOVA main effects, including means and standard deviations, are available in Tables 28-33. From the descriptive statistics in Tables 28-33, it is clear that there were differences in the average percentage of flagged items depending on the pattern of DSF. Except for the Adjacent Category Log Odds Ratio (AC-LOR) statistic, the average percentage of flagged items increased as one goes from "No DSF" to "Divergent" to "Convergent". For example, from Table 30, the average percentage of flagged items from the Generalized Mantel-Haenszel statistic ranged from 5% when No DSF was present, to 14.1% when convergent DSF was present. Also, when going from one DSF pattern to the next, the greatest increase in average percentage of flagged items occurred when going from "No DSF" to "Divergent", except for the Mantel test and Liu-Agresti statistic. In the latter two cases, the greatest increase in average percentage of flagged items occurred when going from the "Divergent" to "Convergent" DSF patterns. For example, from Table 30, the average percentage of flagged items for the Generalized Mantel test with no DSF present was 5% and increased to 13.2% when divergent DSF was present. However, both the Mantel test and and Liu-Agresti statistic started at 5.0% and 4.9% average flagged items with no DSF present, respectively, and both increased to 6.4% when divergent DSF was present.

Table 28. Means and Standard Deviations for Statistical Power using Adjacent Category Log Odds Ratio (AC-LOR)

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	5.0	3.2
Divergent	16000	11.4	5.9
Convergent	16000	9.7	5.1
Item Score Levels			
3 levels	24000	9.1	5.8
4 levels	24000	8.3	5.1
Sample Size Ratio			
600/600	24000	9.2	5.7
1000/200	24000	8.1	5.2
Impact			
No Impact	24000	8.7	5.5
Impact Present	24000	8.7	5.5
Model			
PCM	24000	7.9	5.0
GRM	24000	9.4	5.9

Table 29. Means and Standard Deviations for Statistical Power using Cumulative Log Odds Ratio (CU-LOR)

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	5.0	3.0
Divergent	16000	10.3	4.5
Convergent	16000	11.9	5.2
Item Score Levels			
3 levels	24000	9.1	5.0
4 levels	24000	8.5	4.7
Sample Size Ratio			
600/600	24000	8.6	5.4
1000/200	24000	7.5	4.7
Impact			
No Impact	24000	8.1	5.1
Impact Present	24000	7.9	5.0
Model			
PCM	24000	7.8	4.9
GRM	24000	8.3	5.1

Table 30. Means and Standard Deviations for Statistical Power using the Generalized Mantel Haenszel Test (GMH)

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	5.0	4.0
Divergent	16000	13.2	7.0
Convergent	16000	14.1	7.4
Item Score Levels			
3 levels	24000	10.8	7.2
4 levels	24000	10.8	7.2
Sample Size Ratio			
600/600	24000	11.5	7.5
1000/200	24000	10.1	6.8
Impact			
No Impact	24000	10.8	7.2
Impact Present	24000	10.7	7.2
Model			
PCM	24000	10.4	7.1
GRM	24000	11.1	7.3

Table 31. Means and Standard Deviations for Statistical Power using the Mantel Test

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	5.0	3.9
Divergent	16000	6.4	4.4
Convergent	16000	14.1	7.6
Item Score Levels			
3 levels	24000	8.6	6.5
4 levels	24000	8.4	6.5
Sample Size Ratio			
600/600	24000	9.2	6.9
1000/200	24000	7.8	6.0
Impact			
No Impact	24000	8.5	6.4
Impact Present	24000	8.5	6.5
Model			
PCM	24000	9.2	6.9
GRM	24000	7.8	6.0

Table 32. Means and Standard Deviations for Statistical Power using the Liu-Agresti Statistic

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	4.9	3.9
Divergent	16000	6.4	4.4
Convergent	16000	14.0	7.6
Item Score Levels			
3 levels	24000	8.5	6.5
4 levels	24000	8.4	6.5
Sample Size Ratio			
600/600	24000	9.1	6.9
1000/200	24000	7.7	6.0
Impact			
No Impact	24000	8.5	6.4
Impact Present	24000	8.4	6.5
Model			
PCM	24000	9.2	6.9
GRM	24000	7.7	6.0

Table 33. Means and Standard Deviations for Statistical Power using the Simultaneous Step Level Test

Source	N	Mean	SD
DSF Pattern			
No DSF	16000	11.2	7.6
Divergent	16000	19.0	9.9
Convergent	16000	20.9	10.8
Item Score Levels			
3 levels	24000	17.0	8.6
4 levels	24000	17.1	8.6
Sample Size Ratio			
600/600	24000	17.8	10.8
1000/200	24000	16.3	10.1
Impact			
No Impact	24000	17.0	10.4
Impact Present	24000	17.0	10.5
Model			
PCM	24000	17.2	10.7
GRM	24000	16.9	10.3

CHAPTER 5

DISCUSSION

In this chapter, a synthesis of findings will be discussed. This synthesis will include how findings relate to past literature, revealing which findings agree with or contribute new information to prior research. Finally, limitations will be discussed, as well as further research questions that arise from this project.

Synthesis of Findings

This study considered Type I error and power rates for the Adjacent Category Log Odds Ratio (AC-LOR), Cumulative Category Log Odds Ratio (CU-LOR), Mantel test, Generalized Mantel-Haenszel (GMH) statistic, Liu-Agresti statistic and the Simultaneous Step Level (SSL) test. The a priori Type I error rate was 5% within a +/- 1% margin of error. Statistical adjustments were considered for the SSL test and an ANOVA conducted to determine which of several factors contributed to differences in Type 1 error and power among the DSF and DIF methods.

Type I Error And Power: DSF Detection Methods

Though few studies have been done analyzing and comparing DSF methods, results of this study coincide with results from Penfield (2007, 2008). That is, the AC-LOR statistic has better controlled Type I error and higher power when response data is generated under the partial credit model (PCM) and this is also the case for the CU-LOR statistic when response data is under the graded response model (GRM). Generally, the CU-LOR statistic has higher power than the AC-LOR statistic which makes sense because it utilizes responses from all steps whereas the AC-LOR statistic only utilizes data from adjacent categories. This also means the CU-LOR is more sensitive to DSF pattern as exhibited by some lower power rates in the divergent DSF pattern condition. Higher power also means inflated Type I error for steps that do not have DSF within an item; in this case, the CU-LOR statistic still exhibits better control than the AC-LOR statistic. Although previous studies did not examine other conditions such as sample size ratio or number of score levels, it appears the other factors did not matter as much as the

DSF pattern and after that, the generating model. Penfield (2007) investigated the effect of impact on the CU-LOR and SSL test, where the mean difference between the ability of reference and focal group members was 1, and found slightly raised rejection rates, particularly when item discrimination increased. The present study did not vary item discrimination, had a lower impact value (0.75) and conducted an ANOVA to verify that the effect of impact was not statistically or practically significant (effect size). Having 3 versus 4 score levels did not appear to have an effect on overall performance of the DSF methods, however, false positives may occur in an item that exhibits DSF, but not in all steps, particularly with the AC-LOR statistic.

Type I Error And Power: DIF Detection Methods

Results showed that all Type I error rates were approximately 5% when no DSF was present and when divergent DSF was present; for these DSF patterns, the SSL method approximated a Type I error rate of 5% with statistical adjustments. However, Type I error rate inflation occurred with the convergent DSF pattern, particularly for the Mantel test and Liu-Agresti statistic when items had 3 score levels. These results were expected; Wang and Su (2004) found that the farther the average signed area (ASA) of a test is from zero, the worse Mantel and GMH perform and then once Type I error is lost, Mantel yields larger inflations than GMH. Although Wang & Su (2004) did not vary item score levels, ASA is affected by total number of item score levels. As a reminder, ASA reflects the degree to which the test favors the reference group over the focal group and is found by simply adding all of the signed values of DIF and dividing by the total number of item score levels in the test. Thus, if the magnitude of DSF/DIF stays the same for each item, but the number of item score levels increase from 3 to 4, ASA will automatically decrease and Mantel will have better control over Type I error. For instance, in this study ASA for 3- and 4- item score levels were .075 and .05 respectively. GMH tends to have better control over Type I error than Mantel because the Mantel tests for item means between groups, making it more sensitive to distortion caused by ASA than GMH, which tests overall distributions of score categories.

Under the convergent DSF condition, the Mantel Test seemed to have better control of Type I error under the GRM model, except when the sample size ratio was 1:1. The GMH statistic had better Type I error rates with the GRM when items had 4 score

levels and sample size ratio was 1:1. Wang & Su (2004) made a general statement that Mantel and GMH will have better Type I error control when the PCM is used to generate data due to number correct scores being used as a matching variable which are sufficient statistics and monotonically related to the IRT scales under the PCM. However, Kristjansson, Aylesworth, McDowell, and Zumbo (2005) used the generalized PCM (GPCM) for which raw scores/number correct scores are not sufficient statistics and found no serious Type I error rate problems when raw score DIF detection methods were used. Demars (2008) found that Mantel and GMH methods had similar Type I error rates under the PCM and GRM generated data and suggested that there are factors other than simply the model being used that may cause differences in Type I error rates between the models. In the current study, the Mantel test typically had better Type I error control with the GRM model under convergent DSF. When investigating the conditions that were tested in both studies, a possible reason for the difference in results here versus with Wang & Su (2004) is that the DSF/DIF effects were much larger in the current study. DIF effects in Wang & Su's (2004) study were 0.10 and .25, while the DSF/DIF effects used in the current study were .6, .8, and 1 respectively. Also, Wang & Su (2004) did not investigate the effects of sample size ratio differences between the reference and focal groups on Type I error rates.

Wang & Su (2004) also investigated the effect of impact on Type I error and found that impact only adversely affected Type I error when the difference in mean ability between groups was greater than 1, particularly for the Mantel and GMH under the GRM model. This study did not find a significant effect for impact when the difference in mean ability between groups was 0.75.

The Liu-Agresti statistic is a generalized extension of the Mantel-Haenszel statistic for dichotomous items, estimating the common odds ratio across all K strata for an ordinal response. As a consequence, the Liu-Agresti statistic had very similar behavior as the Mantel test in terms of Type I error rates. Penfield & Algina (2003) showed that the Liu-Agresti statistic had similar rejection rates as Cox's B, where Cox's B is mathematically equivalent to the square root of the Mantel chi square statistic. Thus, it seems to make sense that both the Mantel test and Liu-Agresti statistic had similar behavior in this study.

For the SSL test, Penfield (2007) found that it was more robust than a traditional omnibus DIF detection method when investigating differences in ability distribution, DSF pattern, generating model, and item parameterization. Penfield (2007) did not investigate item score level differences (although it was suggested in future research) nor did he test SSL against several other traditional methods. This current study confirmed that under the Dunn-Bonferroni and Holm adjustments, the SSL test had the best control over Type I error when compared to traditional DIF detection methods. Under the Benjamini-Hochberg adjustment, the SSL test behaved similarly to the GMH statistic. That is, in the convergent DSF condition, more Type I error inflations occured for 3 versus 4 item score levels, particularly when the sample size ratio was 1:1 between the reference and focal groups. In general, it seemed that sample size ratio by itself did not cause differences in Type I error. Although power was slightly lower with the 5:1 sample size ratio, the effect size was fairly small. Wood (2011) found that Type I error for the Mantel test can be controlled with sample size as low as 40. Mantel and Liu-Agresti statistics have been shown to have adequate power (greater than .70) for sample sizes around 250 (Zwick, Donoghue and Grimer, 1993; Zwick and Thayer, 1996). This study confirmed that differences in sample size ratio, by itself, did not adversely affect Type I error or power based on the size of the reference and focal groups for the Mantel and Liu-Agresti statistics in this study and extended these results for the SSL test.

Regarding power, it was not surprising that Mantel had higher power when DIF within the items favored only one group (convergent or constant in other cases) and had poor power when DIF was divergent. This has been well documented (i.e. Zwick, 2012; Zwick and Thayer, 1996; Woods, 2011). Penfield and Algina (2003) confirmed that the Liu-Agresti statistic can be best described as the odds ratio form of the Mantel test, thus it had similar statistical behavior. It has also been well documented that the GMH statistic has much higher power than the Mantel test when DIF is divergent. Information that was not previously known was how the non-traditional DSF method for identifying item DIF, the SSL test, would compare to the traditional DIF methods across various conditions (including number of score levels). The SSL test, though structured differently, had similar behavior to the GMH method. For the Mantel and Liu-Agresti statistic, power rates were slightly lower when items had 4 score levels in the convergent and divergent

DSF pattern under the GRM model, but under the PCM model, power for the 4 item score levels were higher; the opposite pattern was exhibited for both the SSL test and GMH statistic. When looking at single item rejection rates, the SSL test had no inflated Type I error rates in the No DSF condition, while GMH had two inflated rates. With the convergent DSF pattern, there were times when GMH had higher power than SSL, but not consistently. However, when divergent DSF was exhibited, both GMH and the SSL test had similar power rates. Penfield (2009) suggested that, theoretically, DSF methods should not be affected by the number of score levels (although this had not been formally tested) and also encouraged a study comparing the SSL test and Liu-Agresti statistic with other DIF methods. In this study, the pattern of DSF had the most effect on power rates than did other factors, including the number of item score levels.

Impact. It is worth mentioning the results for impact since it did not have as much of an effect on Type I error rates as anticipated. Although, impact is expected to inflate Type I error rates, there actually have been varied results in the literature depending on the magnitude of impact and also other factors that are being considered. For instance, impact has been found to adversely affect Type I error and power for the Mantel and GMH statistic as mean ability difference exceeded 1 (i.e. 1.5) under the GRM model (Wang and Su, 2004). Additionally, when item discrimination increased, slightly raised rejection rates occurred when impact equaled 1 for DSF methods (Penfield, 2007). However, Holland and Thayer (1988) showed that including the studied item ameliorated possible inflated rates caused by impact for the Mantel-Haenszel test, particularly when the underlying model approximates the Rasch model. These previous studies did not indicate an ANOVA model to test the impact effect at different magnitudes of impact. Based on previous research, it is possible that including larger magnitudes of impact, excluding the studied item(s), and/or varying item discrimination in the GRM model would have contributed to seeing a larger effect for impact. However, in this study, item discrimination was equal to 1 and not varied, studied items were included in the observed test score, impact was less than 1, and utilizing an ANOVA showed that impact did not have a large effect when compared with other factors used in this study (eta was less than 0.1). The pattern of DSF superseded the effect of any other factor in this study.

Statistical Adjustments

Although Kim (2010) suggested that DIF detection methods may suffer Type I error inflation due to the testing of multiple items for DIF in an exam, this study did not confirm that to be an issue. Wang and Su (2004) suggested that inflated Type I error is not as much of a problem for polytomous item DIF detection which was found to be true in this study. Inflated rates were due to changes in DSF pattern, particularly if the convergent DSF pattern was present within an item. A reason for using statistical adjustments is if multiple tests are being compared in order to make a decision on an item, which is exactly the case for the SSL test. The SSL test uses a statistical test at each step of an item to determine overall DIF. Penfield (2007) made adjustments at each step of the item, while this study made the adjustment after all steps had been tested for DSF. While Penfield (2007) used the popular Dunn-Bonferroni adjustment, he recommended other adjustments should be investigated as well. What Works Clearinghouse (2014) recommended the Benjamini-Hochberg adjustment because it controls Type I error but still leaves adequate power for detecting DIF. In this study, results from the Holm adjustment was very similar to Dunn-Bonferroni, the Benjamini-Hochberg was slightly less conservative than Dunn-Bonferroni, but not by much. If Type I error was very highly inflated, there may have been more variation among the different statistical adjustments; however the Dunn-Bonferroni adjustment was appropriate to use for this study.

Factors Affecting Power: ANOVA Study

In order to identify the most important effect(s) in the ANOVA , eta squared was used as an effect size to accompany *p*-values. When combining effect size and *p*-value, it was found that the DSF pattern was the most important effect, as compared to other factors, for the DSF/DIF methods in this study. Wood (2011) utilized an ANOVA for several DIF methods, including the Mantel test and Liu-Agresti and also found a significant effect for DIF pattern as well as other effects that were not used in this study. An advantage of nonparametric DIF detection methods is that they are not as sensitive to generating model, differences in population distributions or smaller sample sizes as their parametric counter parts. Although in this study generating model, sample size ratio and impact did cause slight changes in rates, the ANOVA study was helpful in showing that

these effects were not large enough to make a difference, particularly in comparison to the DSF pattern.

Limitations and Future Research

Due to the many conditions that would result, not every scenario could be investigated. With 48 conditions being investigated, several factors were used in the ANOVA study, leading to many interaction effects; however, effect size values were fairly low once the DSF pattern was accounted for. Since the Liu-Agresti statistic had such similar behavior to the Mantel test, another method could have been used for comparison. Additionally, the SSL test could be compared to parametric methods for identifying DSF. One of the benefits of DSF methods is that they give an effect size of the item DSF along with statistical significance. The accuracy of the DSF methods for recovering effect sizes under various conditions could be investigated in a future study. Additionally, one could determine if the average of DSF effects for each item are equal to the net item DIF effect found by such methods as Liu-Agresti and Cox Beta. It was suggested by Gattamorta, Penfield, & Myers (2012) that this equality should hold true but had only been tested using the IRT PCM model to detect DSF.

This study compared several DSF/DIF detection methods to determine which had better control over Type I error and higher power rates, what statistical adjustments were best to use under multiple significance testing for DIF, and determined what factors affect power rates in these methods as item score levels change. For practical purposes, results confirmed that changing the structure of an item by adding or removing score levels did not have an effect on the DSF/DIF detection methods used in this study. Changes to the number of items flagged for DIF that may occur if the number of score levels change is most likely due to how the introduction of the score level affected the pattern of DIF within the item, which essentially means that the addition or removal of a score level is only a problem if it adversely affects the magnitude of DIF in the item. As for the specific DSF and DIF methods, the CU-LOR statistic was more stable than the AC-LOR statistic; additionally, the SSL test performed similarly to the GMH statistic and both were better than Mantel and Liu-Agresti when considering both convergent and divergent DSF. The average signed area (ASA) will be lower as items increase in number of score

levels without the magnitude of DIF increasing. In this case, polytomous DIF detection methods will have fairly well controlled Type I error rates. In this study, inflation rates of the SSL test were adequately adjusted using the Dunn-Bonferroni adjustment. Finally, when considering both statistical significance and effect size of the factors affecting power in this study (DSF pattern, sample size ratio, generating model, impact and number of item score levels), the pattern of DSF was most important for the DIF/DSF detection methods investigated. Due to how well the SSL test performed when compared to the other methods in this study, it should be considered as a way to comprehensively analyze items for DIF, as it allows one to simultaneously investigate both DSF and DIF in a robust way under various conditions.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ankenmann, R. D., Witt, E. A., and Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300. Doi: 10.111/j.1745-3984.tb00558.x

Atar, B. (2007). *Differential item functioning analysis for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures.* (Doctoral dissertation). The Florida State University Department of Educational Psychology and Learning Systems.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*(4), 1165-1188.

Bolboaca, S. D., Jantschi, L., Sestras, A. F., Sestras, R. E., & Panfil, D. C. (2011). Pearson-Fisher Chi-Square Statistic Revisited. *Information, 2*, 528-545. Doi: 10.3390/info2030528

Bolt, D. M. & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 43*(4), 313-333. Doi: 10.1111/j.1745-3984.2006.00019.x

Bonferroni, C. E. (1936). Statistical class theory and calculation of probability. *Publication of High R Institute of Economic and Commerical Sciences of Florence, N.8.*).

Camilli, G. (2006). Test fairness. In R.L. Brennan (Ed.), *Educational measurement* (4th

ed.) (pp. 221-256). Westport, CT: American Council on Education and Praeger Publishers.

Camilli, G. & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24,* 323-341.

Camilli, G. & Shepard, L.A. (1994). Methods for identifying biased test items. Hollywood, CA: Sage Publications.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Chang, H.-H. & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391-404.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*(1), 67-78.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269-279.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*(1), 67-78.

Cohen (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum

Cohen, A. S. & Kim, S. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*, 39-52.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B, 20*(2), 215-242.

De Ayala, R.J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 25,* 172-189.

De Ayala, R.J. (2009). *The theory and practice of Item Response Theory.* New York: The Guilford Press.

DeMars, C.E. (2008). Polytomous DIF and violations of ordering of the expected latent trait by the raw score. *Educational and Psychological Measurement, 68,* 379-396.

Donoghue, J. R. & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*(2), 131-154.

Dorans, N. & Holland, P. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35-35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Dorans, N. J. & Schmitt, A. P. (1991). *Constructed-response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-47). Princeton, NJ: Educational Testing Service. [Also appears in *Construction vs. Choice in Cognitive Measurement* (pp. 135-166), R. E. Bennett & W. C. Ward (Eds.), 1993, Hillsdale, NJ: Erlbaum.]

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134 – 135.

Drasgow, F (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19 – 29.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *Journal of Experimental Education, 73*, 23-39.

Fidalgo, A. M. & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for DIF detection *Educational and Psychological Measurement, 68,* 940-958.

Fidalgo, A. M., Mellenbergh, G. J., & Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online 5*(3), 43-53.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muniz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education, 75,* 293-314.

Finch, W. H. & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. Educational and Psychological Measurement, *67,* 565-582.

Fleming, K., Ross, M., Tollefson, N., & Green, S. B. (1998). Teacher's choices of test-item formats for classes with diverse achievement levels. *Journal of Educational Research, 91*(4), 222-228.

Gattamorta, K.A. & Penfield, R.D. (2012). A comparison of adjacent categories and cumulative differential step functioning estimators. *Applied Measurement in Education, 25*(2)*, 142-161.*

Gattamorta, K.A., Penfield, R.D., & Meyers (2012). Modeling item-level and step-level invariance effects in polytomous items using the partial credit model. *International Journal of Testing, 12*(3)*, 252-272.*

Gilmore, W. (2014). *A differential item functioning analysis of the new Mexico English Language Proficiency Assessment.* (Doctoral Dissertation). Retrieved from http://hdl.handle.net/1928/24311.

Guilera, G., Gomez-Benito, J., Hidalgo, M. D., & Sanchez-Meca, J. (2013). Type I error and statistical power of Haenszel procedure for detecting DIF. A meta-analysis. *Psychological Methods*. Doi: 10.1037/a0034306

Hambleton, R. K. & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 3,* 244-253.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66,* 487-506.

Hidalgo, M. D. & Gomez, J. (2006). Nonuniform DIF detection using discriminant logistic analysis and multinomial logistic regression: a comparison for polytomous items. *Qualtitative Quantitative International Journal of Methodology, 40*(5), 805-823.

Hidalgo-Montesinos, M. D. & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62*(1), 32-44.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383-386.

Holland, P. W. & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Kim, J. (2010). Controlling Type I error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjusting multiple item testing. (Doctoral dissertation). *Educational Policy Studies Dissertations. Paper 67.*

Kim, S., Cohen, A. S., & Kim, H. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*, 217-228.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65,* 935-953.

Kabasakal, K. A., Arsan, N., Gok, B., & Kelecioglu, H. (2014). Comparing performances (type I error and power) of IRT likelihood ratio, SIBTEST, and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory & Practice, 14*(6), 2175-2193.

Kaya, Y., Leite, W. L., & Miller, M. D. (2015). A comparison of logistic regression models for DIF detection in polytomous items: the effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education, 2*(1), 22-39.

Kristjansson, E., Aylesworth, R. McDowell, I., & Zumbo, B.D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement,* 65, 935-953.

Kwak, N., Davenport, E. C., Jr., & Davison, M. L. (1998). *A Comparative Study of Observed Score Approaches and Purification Procedures for Detecting Differential Item Functioning.* Paper presented at the Annual Meeting of the National Council on

Measurement in Education, April 1998, San Diego, CA.

Li, H. H. & Stout, W. F. (1996). A new procedure for detection of crossing DIF. *Pschometrika, 61,* 647-677.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erbaum Associates.

Liu, I. M. & Agresti, A. (1996). Mantel-Haenszel-Type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52* 1223-1234.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58, 690-700.*

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute 22,* 719-748.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Mellenbergh, G. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118. doi: 10.2307/1164960

Meredith, W. & Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311.

Meulders, M. & Xie, Y. (2004). *Explanatory item response models* (p. 213-240). Springer: New York.

Miller, R. G. (1981) *Simultaneous statistical inference* (2nd edition). Springer: Verlag.

Miller, T., Chanine, S., & Childs, R. A. (2010). Detecting differential item functioning and differential step functioning due to differences that should matter. *Practical Assessment, Research & Evaluation, 15*(10), 1-12. Available online: http://pareonline.net/getvn.asp?v=15&n=10.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334. doi:

10.1177/014662169301700401

Miles J. & Shevlin M. (2001). *Applying regression and correlation: A Guide for Students and Researchers.* Sage: London.

Monahan, P. O. & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on Type I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement,* 42(2), 101-131.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies detect flaws in item translations. *International Journal of Testing, 1,* 115-135.

Nandakumar, R. (2005). Assessing dimensionality of a set of item responses-comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Narayanan, P. & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-338.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Navas-Ara, M. J. & Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment 18*(1), 9-15.

Park, D. G. & Lautenschlager, G. J. (1990). Iterative linking and ability scale purification as means for improving IRT item bias detection. *Applied Psychological Measurement, 14,* 163-173.

Penfield (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement Education 14*(3): 235-259.

Penfield, R. D. & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15.

Penfield, R. D. & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of*

Educational Measurement, 40(4), 353-370.

Penfield, R. D. (2007). Assessing differential functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3)*, 187-210.*

Penfield, R. D. (2008). Three classes of nonparametric differential step functioning estimators. *Applied Psychological Measurement, 32*(6)*, 480-501.*

Penfield, R. D. & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C.r. Rao (Eds.) *Handbook of Statistics, Volume 26: Psychometrics* (pp. 125-167). New York: North Holland.

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 25*(1)*,* 61-78.

Penfield, R. D. (2010). Explaining crossing DIF in polytomous items using differential step functioning effects. *Applied Psychological Measurement, 34*(8), 563-579. doi:10.1177/0146621610377083

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-27. doi: 10.1177/014662169501900104.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197-207.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33,* 135-147.

Raju, N., Laffitte, L., & Byrne, B. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Reise, S., Widaman, K., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance.

Psychological Bulletin, 114(3), 552.

Roussos, L. A. & Stout, W. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Richmond, Va.: Psychometric Society.

Schulz, W. & Fraillon J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6), 447-464.

Shealy, R. & Stout, W. F. (1991). *An item response theory model for test bias.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Smith, R. (2004). Detecting item bias with the Rasch model. *Journal of Applied Measurement, 5*(4): 430-449.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40,* 106-108.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*, 332-342.

Steinberg, L. & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402-415. doi: 10.1037/1082-989X.11.4.402.

Stone, B. J. (1992). Joint confirmatory factory analyses of the DAS and WISC-R. *Journal of School Psychology, 30*, 185-195.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using

logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Su, Y. & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning in polytomous items. *Applied Measurement in Education, 18,* 313-350.

Taylor, C. S. & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25,* 246-280.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77-83.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Wang, W. & Su, Y. (2004). Influencing the Mantel and Generalized Mantel-Haenszel method for the assessment of differential item functioning in polytomous items.

Welch, C. J., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1-19.

Williams, V., Jones, L., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state to state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24,* 42-69.

Wood, S.W. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small.* (Doctoral dissertation). Retrieved from Proquest Dissertation and Theses. (Access Order No. UMI 346258).

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32*, 511-526.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item

functioning. *Applied Psychological Measurement, 33,* 42-57.

Yamamoto, K. & Muraki, E. (1991). *Non-linear transformation of IRT scale to account for the effect of non-normal ability distribution on item parameter estimation.* A paper presented at the annual 1991 American Educational Research Association meeting, Chicago, IL.

Yanagawa, T. & Fujii, Y. (1990). Homogeneity test with a generalized Mantel-Haenszel estimator for L2 x K contingency tables. *Journal of the American Statistical Association, 85*, 744-748.

Zeiky, M. (1993). Practical question in the use of DIF statistics in item development. In P. W. Holland & Wainer (Eds.), *Differential item functioning* (pp. 337-347), Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational and Behavioral Statistics, 15,* 185-197.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report. No. RR-12-08). Princeton, NJ: Educational Testing Service.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R. & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item. *Applied Psychological Measurement, 26*(1), 57-76.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21,* 187-201.

Zwick, R., Thayer, D.T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121-140.

Zwinderman, A. H. & Van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurment, 14*, 73-81. doi: 10.1177/014662169001400107.

APPENDIX A: EXAMPLE MULTIPLE SCORE LEVEL ITEM

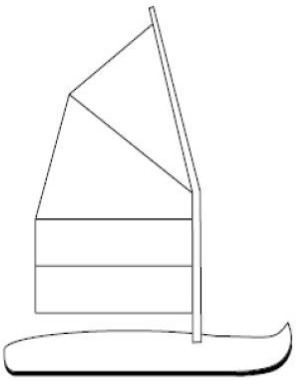
Practical Assessment, Research & Evaluation, Vol 15, No 10
Miller, Chahine, & Childs, DIF and DSF

Appendix B

Item P34 and scoring guide (in *Sample Assessment Questions and Scoring Guides*, available at www.eqao.com).

Marco is making the sail using green and red material in the ratio 3:2. He needs a total of 4.5 m^2 of material.

d) Determine how much **red** material he needs.
Show your work.



d)	10	Application of knowledge and skills to determine the amount of red material, using ratios, shows limited effectiveness due to • misunderstanding of concepts • incorrect selection or misuse of procedures (e.g., does not multiply or divide, or uses numbers other than 2, 3, 5)
	20	Application of knowledge and skills to determine the amount of red material, using ratios, shows some effectiveness due to • partial understanding of the concepts • errors and/or omissions in the application of the procedures (e.g., multiplies or divides with one of 2, 3, 5)
	30	Application of knowledge and skills to determine the amount of red material, using ratios, shows considerable effectiveness due to • an understanding of most of the concepts • minor errors and/or omissions in the application of the procedures (e.g., multiplies and divides by wrong numbers [2, 3, 5])
	40	Application of knowledge and skills to determine the amount of red material, using ratios, shows a high degree of effectiveness due to • a thorough understanding of the concepts • an accurate application of the procedures (any minor errors and/or omissions do not detract from the demonstration of a thorough understanding) (e.g., multiplies by 2 and divides by 5 to get 1.8 m^2)

APPENDIX B: P-VALUE ADJUSTMENT GRAPHS

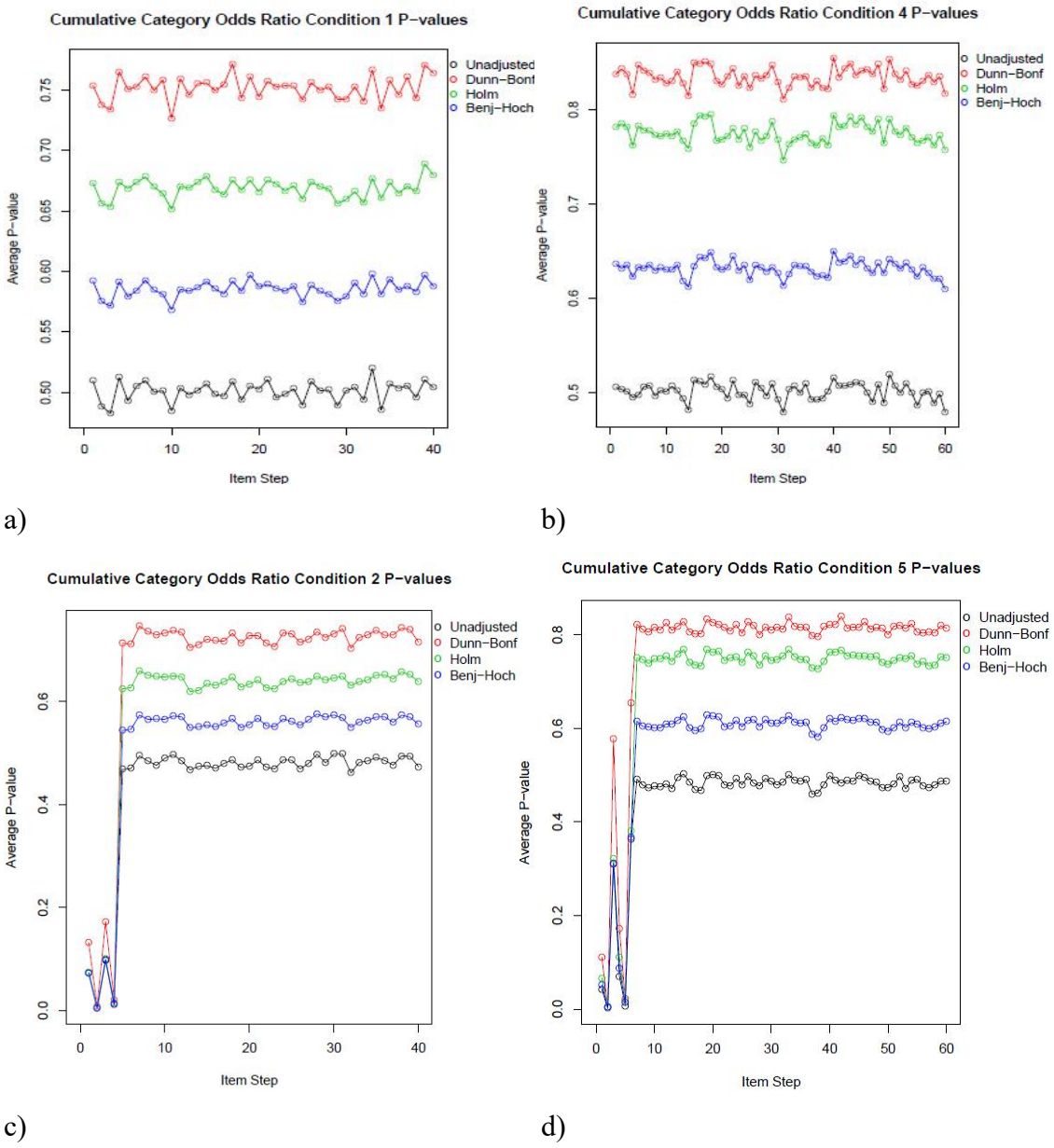
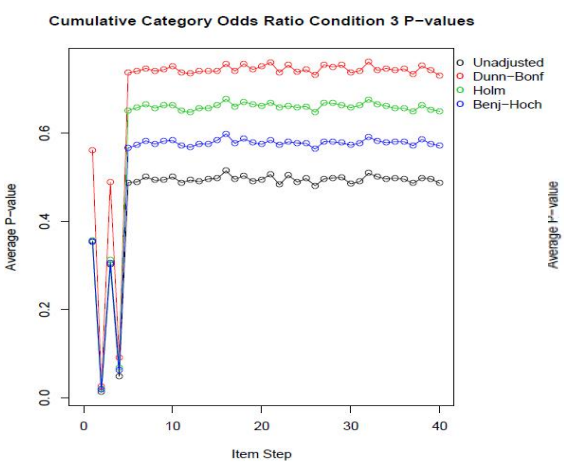
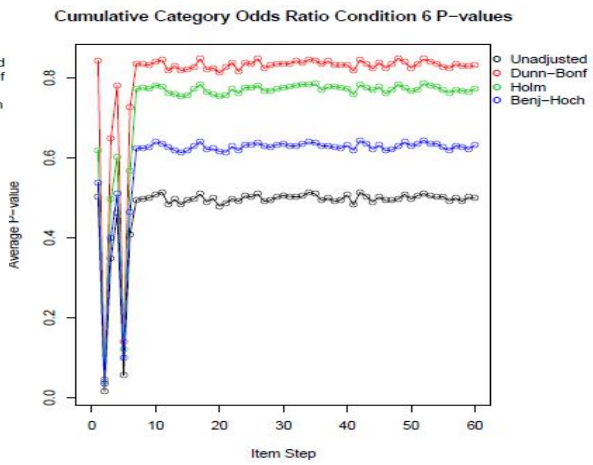


Figure 2 . Cumulative category log odds ratio (CU-LOR) *p*-values for selected conditions



e)



f)

Figure 2. (cont'd) Cumulative category log odds ratio (CU-LOR) *p*-values for selected conditions*

*Conditions:

1	pcm	No impact	600/600	3 levels	No DSF
2	pcm	No impact	600/600	3 levels	convergent
3	pcm	No impact	600/600	3 levels	divergent
4	pcm	No impact	600/600	4 levels	No DSF
5	pcm	No impact	600/600	4 levels	convergent
6	pcm	No impact	600/600	4 levels	divergent

APPENDIX C: R FUNCTIONS FOR DATA GENERATION AND CREATION OF DIF/DSF DETECTION METHODS

```
#####################################################

#### Program for Creating Batchfiles for WinGen ####

#####################################################


#### Model parameters ####

ratio.1 <- c(rep(600,6), rep(1000,6))

ratio.r <- c(rep(ratio.1,4))

ratio.2 <- c(rep(600,6), rep(200,6))

ratio.f <- c(rep(ratio.2,4))

sample.size <-  c(ratio.r,ratio.f)


means.r <- c(rep(0,48))

means.2 <- c(rep(0,12), rep(-0.75,12))

means.f <- c(rep(means.2,2))

impact <- c(means.r,means.f)


#### Name of item files ####

PRfiles<- c("PcmRNO3.wgi", "PcmRCONS3.wgi", "PcmRDIV3.wgi",

            "PcmRNO4.wgi", "PcmRCONS4.wgi", "PcmRDIV4.wgi")
```

```
GRfiles<- c("GrmRNO3.wgi", "GrmRCONS3.wgi", "GrmRDIV3.wgi",
            "GrmRNO4.wgi", "GrmRCONS4.wgi", "GrmRDIV4.wgi")


Ritem.files <- c((rep(PRfiles,4)),(rep(GRfiles,4)))

PFfiles <- c("PcmFNO3.wgi", "PcmFCONS3.wgi", "PcmFDIV3.wgi",
            "PcmFNO4.wgi", "PcmFCONS4.wgi", "PcmFDIV4.wgi")

GFfiles<-   c("GrmFNO3.wgi", "GrmFCONS3.wgi", "GrmFDIV3.wgi",
            "GrmFNO4.wgi", "GrmFCONS4.wgi", "GrmFDIV4.wgi")

Fitem.files<- c((rep(PFfiles,4)),(rep(GFfiles,4)))
item.files <- c(Ritem.files, Fitem.files)

#### Model types ####
model <- c(rep("PCM", 24),rep("GRM", 24),rep("PCM",
24),rep("GRM", 24))

#### Loop through Conditions ####
for(i in 1:96){
```

```
  #### Create batchfiles ####
  if(i <= 48){

sink(paste("F:\\Alicia_Drive\\EdPsy\\PhDProgram\\Dissertatio
n\\Prospectus\\Software_Resources\\WinGen\\batchfiles\\condi
tion",i, "r.wgs", sep=""))
  }
  else if(i>48){

sink(paste("F:\\Alicia_Drive\\EdPsy\\PhDProgram\\Dissertatio
n\\Prospectus\\Software_Resources\\WinGen\\batchfiles\\condi
tion",(i-48), "f.wgs", sep=""))
  }
  cat(paste(sample.size[i]))
  cat(",normal,")
  cat(paste(impact[i]))
  cat(",1")
  cat("\n")

  #### Input item files ####

cat(paste("file,F:\\Alicia_Drive\\EdPsy\\PhDProgram\\Dissert
```

```
ation\\Prospectus\\Software_Resources\\WinGen\\itemfiles\\",
item.files[i],sep="" ))
  cat("\n")


  #### Output response files ####
  if(i <= 48){

cat(paste("C:\\Users\\Michael\\Documents\\wingen_files\\cond
ition",i, "r.wgr", sep=""))
  }
  else if(i>48){

cat(paste("C:\\Users\\Michael\\Documents\\wingen_files\\cond
ition",(i-48), "f.wgr", sep=""))
  }
  cat("\n")
  cat("replicate,1000")
  cat("\n")
  cat(paste(model[i]))
  cat("\n")
  cat("random")
  cat("\n")
  cat("none")
```

```
  cat("\n")
  cat("none")

  sink()

}

##########################
## Cuefile for Win Gen ##
##########################

sink("F:\\Alicia_Drive\\EdPsy\\PhDProgram\\Dissertation\\Pro
spectus\\Software_Resources\\WinGen\\batchfiles\\cuefile.wgc
")
for(i in 1:96){
  if(i <= 48){

cat(paste("F:\\Alicia_Drive\\EdPsy\\PhDProgram\\Dissertation
\\Prospectus\\Software_Resources\\WinGen\\batchfiles\\condit
ion",i, "r.wgs", sep=""))
    cat("\n")
  }
```

```
  if(i >
48){    cat(paste("F:\\Alicia_Drive\\EdPsy\\PhDProgram\\Diss
ertation\\Prospectus\\Software_Resources\\WinGen\\batchfiles
\\condition",(i-48), "f.wgs", sep=""))

    cat("\n")

  }

}

sink()

#####################################

#Adjacent category odds ratio procedure
#####################################
AC.OR <- function(table.data){
    stratum <- dim(table.data)[3]
    steps <-(dim(table.data)[2])-1

    #Creating all necessary vectors/matrices
    A.jk <- matrix(c(0), nrow=steps,ncol=stratum) #added
vectors A-D
    B.jk <- matrix(c(0), nrow=steps,ncol=stratum)
    C.jk <- matrix(c(0), nrow=steps,ncol=stratum)
    D.jk <- matrix(c(0), nrow=steps,ncol=stratum)
    N.jk <- matrix(c(0), nrow=steps,ncol=stratum)
    se.ljnum <- matrix(c(0), nrow=steps,ncol=stratum)
    se.ljdenom <- matrix(c(0), nrow=steps,ncol=stratum)
    se.lj <- rep(0, times=steps)

    alpha.jnum <- matrix(c(0), nrow=steps,ncol=stratum)
    alpha.jdenom <- matrix(c(0), nrow=steps,ncol=stratum)
    alpha.j <- rep(0, times=steps)
```

```
    alpha.vector <- rep(0, times=stratum)

      for(j in 1:steps){

      for(k in 1:stratum) {
        #parts of odds ratio being created
        A.jk[j,k] <- table.data[1,(j+1),k] #for the
reference group
        B.jk[j,k] <- table.data[1,j,k]
        C.jk[j,k] <- table.data[2,(j+1),k] #for the focal
group
        D.jk[j,k] <- table.data[2,j,k]

        #total respondents
        N.jk[j,k] <- (A.jk[j,k] + B.jk[j,k] + C.jk[j,k] +
D.jk[j,k])

        #odds ratio numerator and denominator
        alpha.jnum[j,k] <- (A.jk[j,k]*D.jk[j,k])/N.jk[j,k]
        alpha.jdenom[j,k] <- (B.jk[j,k]*C.jk[j,k])/N.jk[j,k]

        #removing NANs
        alpha.jnum[!is.finite(alpha.jnum)] <- 0
        alpha.jdenom[!is.finite(alpha.jdenom)] <- 0

      }
    }

    for(j in 1:steps){
      #odds ratio
      alpha.j[j] <- sum(alpha.jnum[j,])/sum(alpha.jdenom[j,])
```

```
    }

    #log-odds ratio

    lambda.j <- log(alpha.j)

    # Standard error
    for(j in 1:steps){

      for(k in 1:stratum) {
        se.ljnum[j,k] <- (N.jk[j,k])^(-
2)*(A.jk[j,k]*D.jk[j,k]+alpha.j[j]*B.jk[j,k]*C.jk[j,k])*

(A.jk[j,k]+D.jk[j,k]+alpha.j[j]*B.jk[j,k]+alpha.j[j]*C.jk[j,
k])

        se.ljdenom[j,k] <- (A.jk[j,k]*D.jk[j,k])/(N.jk[j,k])

        se.ljnum[!is.finite(se.ljnum)] <- 0
        se.ljdenom[!is.finite(se.ljdenom)] <- 0
      }
    }

    for(j in 1:steps){
      se.lj[j] <-
sqrt(sum(se.ljnum[j,])/(2*(sum(se.ljdenom[j,]))^2))
    }

    #Z test statistic
    z.stat <- lambda.j/se.lj

    #p-value
```

```
    pval <- 2*pnorm(-abs(z.stat))


return(list(pval=pval,lambda.j=lambda.j,se.lj=se.lj,z.stat=z.
stat))

  #} #for-loop close
}#ACOR function close
```

```
################################
##Cumulative category odds ratio
################################
CU.OR <- function(table.data){

  #table.data <- item.table
  stratum <- dim(table.data)[3]
  steps <-(dim(table.data)[2])-1

  #Creating all necessary vectors/matrices
  A.jk <- matrix(c(0), nrow=steps,ncol=stratum) #added
vectors A-D
  B.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  C.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  D.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  N.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  se.ljnum <- matrix(c(0), nrow=steps,ncol=stratum)
  se.ljdenom <- matrix(c(0), nrow=steps,ncol=stratum)
  se.lj <- rep(0, times=steps)

  alpha.jnum <- matrix(c(0), nrow=steps,ncol=stratum)
  alpha.jdenom <- matrix(c(0), nrow=steps,ncol=stratum)
  alpha.j <- rep(0, times=steps)


  alpha.vector <- rep(0, times=stratum)


  #cu.or <- function(test.data, item.scores) {
  for(j in 1:steps){

    for(k in 1:stratum) {
```

```
      #parts of odds ratio being created
      A.jk[j,k] <- sum(table.data[1,c((j+1):(steps+1)),k])
      B.jk[j,k] <- sum(table.data[1,c(1:j),k])
      C.jk[j,k] <- sum(table.data[2,c((j+1):(steps+1)),k])
      D.jk[j,k] <- sum(table.data[2,c(1:j),k])

      #total respondents
      N.jk[j,k] <- (A.jk[j,k] + B.jk[j,k] + C.jk[j,k] +
D.jk[j,k])

      #odds ratio numerator and denominator
      alpha.jnum[j,k] <- (A.jk[j,k]*D.jk[j,k])/N.jk[j,k]
      alpha.jdenom[j,k] <- (B.jk[j,k]*C.jk[j,k])/N.jk[j,k]

      alpha.jnum[!is.finite(alpha.jnum)] <- 0
      alpha.jdenom[!is.finite(alpha.jdenom)] <- 0
    }
  }
  for(j in 1:steps){
    #odds ratio
    alpha.j[j] <- sum(alpha.jnum[j,])/sum(alpha.jdenom[j,])
  }
  #log odds ratio
  lambda.j <- log(alpha.j)

  # Standard error
  for(j in 1:steps){

    for(k in 1:stratum) {
      se.ljnum[j,k] <- (N.jk[j,k])^(-
2)*(A.jk[j,k]*D.jk[j,k]+alpha.j[j]*B.jk[j,k]*C.jk[j,k])*
```

```
(A.jk[j,k]+D.jk[j,k]+alpha.j[j]*B.jk[j,k]+alpha.j[j]*C.jk[j,
k])

      se.ljdenom[j,k] <- (A.jk[j,k]*D.jk[j,k])/(N.jk[j,k])

      se.ljnum[!is.finite(se.ljnum)] <- 0
      se.ljdenom[!is.finite(se.ljdenom)] <- 0
    }
  }

  for(j in 1:steps){
    se.lj[j] <-
sqrt(sum(se.ljnum[j,])/(2*(sum(se.ljdenom[j,]))^2))
  }

  #Z test statistic
  z.stat <- lambda.j/se.lj
  pval <- 2*pnorm(-abs(z.stat))
return(list(pval=pval,lambda.j=lambda.j,se.lj=se.lj,z.stat=z.
stat))

} #close function
```

```
################################################
#Simultaneous Step Level (SSL) DIF Test Function
################################################
#Input: test.data: raw test data
#       p.values: Cumulative log odds ratio (CU-OR) p-values
for test item steps (DSF)
#Output: ssl.vector: Vector indicating item DIF under SSL
for each item
#Completed: 8/27/2014

SSL <- function(step.value,p.values){

  temp <- rep(0, times=step.value) #holds temporary DSF
indicators for each item step

 for(i in 1:step.value){
    ifelse(p.values[i] < 0.05,temp[i]<- 1,temp[i] <- 0)
#indicates DIF by using item DSF p-values
    }
  ifelse(sum(temp)>0, ssl.indicator <- 1, ssl.indicator <- 0)
#DIF indication per item
  return(list(ssl.indicator=ssl.indicator)) #p-values for
this vector are based on DSF p-values previously calculated
}
################################################
#Adjusted Simultaneous Step Level (SSL) DIF Test Function
(Uses adjusted p-values from CU-LOR function)
################################################
#Input: test.data: raw test data
#       p.values: Cumulative log odds ratio (CU-OR) p-values
for test item steps (DSF)
#Output: ssl.vector: Vector indicating item DIF under SSL
```

```
for each item
#Completed: 8/27/2014

adjust.SSL <- function(items,step.value,p.values){
  adjust.ssl.vector <- rep(0, times=items)  #Vector to
indicate item DIF (1=yes/0=no) under SSL
  temp <- rep(0, times=(items*step.value)) #holds temporary
DSF indicators for each item step

 k <-1 #counter for ssl.vector
 for(i in seq(1,(items*step.value), by=step.value)){

   for(j in 1:(items*step.value)){
    ifelse(p.values[j] < 0.05,temp[j]<- 1,temp[j] <- 0)}
#indicates DIF by using item DSF p-values

  ifelse(sum(temp[i:(i+step.value-1)])>0,
adjust.ssl.vector[k] <- 1, adjust.ssl.vector[k] <- 0)
  #DIF indication per item

  k <- k +1

}

return(list(adjust.ssl.vector=adjust.ssl.vector)) #p-values
for this vector are based on DSF p-values previously
calculated

}
```

```
###################################
#Mantel Test for polytomous items
###################################
#Input: table.data..2 x J x K table for the studied item
#       item.scores: potential scores of the studied item
#Output: X2.value: chi-square (Mantel) statistic
#        p.value: p-value for the Mantel hypothesis test
#Notation follows from Wang & Su (2004)
#Completed 8/27/2014

Mantel <- function(table.data){

  group   <- dim(table.data)[1]   #groups
  level.yT  <- dim(table.data)[2] #categories/score levels
  stratum.k <- dim(table.data)[3] #stratum
  item.scores <- as.numeric(colnames(table.data)) #item
scores

  #Assigning variable vectors for contingency table and test
statistic computations
  Fk <- rep(0, times=stratum.k)
  EFk <- rep(0, times=stratum.k)
  VFk <- rep(0, times=stratum.k)
```

```
  n.plusjk <- rep(0, times=level.yT)

  for(k in 1:stratum.k) {
    Fk[k] <- sum(item.scores*table.data[2,,k])  #sum of
focal group scores at kth level

    for(j in 1:level.yT){
      n.plusjk[j] <- sum(table.data[,j,k])} #sum of freqs at
each item score level

    n.Rplusk <- sum(table.data[1,,k])  #sum of reference
group frequencies at each item score level
    n.Fplusk <- sum(table.data[2,,k])  #sum of focal group
frequencies at each item score level
    n.plusplusk <- sum(table.data[,,k]) #total sum of
frequencies

    EFk[k] <-
(n.Fplusk/n.plusplusk)*sum(item.scores*n.plusjk) #expected
value of Fk

    #Pieces used to calculate variance
    piece1 <- (n.Rplusk * n.Fplusk)/((n.plusplusk^2) *
(n.plusplusk-1))
    piece2 <- n.plusplusk * sum((item.scores^2) * n.plusjk)
    piece3 <- sum(item.scores * n.plusjk)

    #clearing NANs
    piece1[!is.finite(piece1)] <- 0
    piece2[!is.finite(piece2)] <- 0
    piece3[!is.finite(piece3)] <- 0
```

```
    VFk[k] <- piece1 * (piece2-(piece3^2)) #Variance of Fk
  }

  X2.value <- ((sum(Fk) - sum(EFk))^2)/sum(VFk) #Chi-square
test statistic
  p.value <- pchisq(X2.value, df=1, lower.tail = FALSE) #p-
value

  return(list(X2.value=X2.value,p.value=p.value))
}
```

```
##############################################
#Generalized Mantel Haenszel (GMH) Function
##############################################

#Input: 2 x group x strata contingency table for studied
item
#Output: GMH Chi-square test statistic and p-value
#Notation follows from Fidalgo (2008)
#Completed: 8/26/2014

GMH <- function(table.data){
```

```
groups.R   <- dim(table.data)[1] #groups
cats.C  <- dim(table.data)[2]    #categories/score levels
strata.h <- dim(table.data)[3]  #strata

#Assigning variable vectors/matrices/arrays for contingency
table computations
n.h <- matrix(c(0), nrow=(groups.R*cats.C), ncol=strata.h)
m.h <- matrix(c(0), nrow=(groups.R*cats.C), ncol=strata.h)
V.h <- array(dim=c(groups.R*cats.C, groups.R*cats.C,
strata.h))
D.ph.x <- array(dim=c(cats.C, cats.C, strata.h))
D.phx. <- array(dim=c(groups.R, groups.R, strata.h))
N.h.. <- rep(0, times=strata.h)
phx. <- matrix(c(0), nrow=groups.R, ncol=strata.h)
ph.x <- matrix(c(0), nrow=cats.C, ncol=strata.h)

#Assigning variable vectors/matrices for linear functions
used based on hypothesis test
I.rminus1 <- matrix(diag(1, (groups.R-1), (groups.R-1)),
(groups.R-1), (groups.R-1))
J.rminus1 <- matrix(c(1), groups.R-1, 1)
I.cminus1 <- matrix(diag(1, (cats.C-1), (cats.C-1)),(cats.C-
1),(cats.C-1))
J.cminus1 <- matrix(c(1), cats.C-1, 1)
R.h <- matrix(c(I.rminus1, -J.rminus1), groups.R-1, groups.R)
C.h <- matrix(c(I.cminus1, -J.cminus1), cats.C-1, cats.C)
A.h <- kronecker(C.h,R.h)

##Assigning variable matrices/arrays for parts of test
statistic
piece1 <- matrix(c(0), strata.h, ((groups.R-1)*(cats.C-1)))
piece2 <- array(dim=c(((groups.R-1)*(cats.C-1)), ((groups.R-
```

```
1)*(cats.C-1)), strata.h))
piece3 <- matrix(c(0), ((groups.R-1)*(cats.C-1)), strata.h)

x <- 1  #counter for observed frequencies

for(k in 1:strata.h) {
  N.h..[k] <- sum(table.data[,,k]) #total number of
responses

  for(j in 1:cats.C){
    n.h[c(x:(x+1)),k] <- table.data[1:groups.R,j,k]
#observed frequences
    ph.x[j,k] <- (sum(table.data[,j,k]))/N.h..[k]
#expected column proportions

    ifelse(x < ((groups.R*cats.C)-1), x <- x+2, x <- 1)}

    for(i in 1:groups.R){
      phx.[i,k] <- (sum(table.data[i,,k]))/N.h..[k]}
#expected row proportions

  m.h[,k] <- N.h..[k]*kronecker(ph.x[,k],phx.[,k])
#expected frequencies
  D.ph.x[,,k] <- diag(ph.x[,k])
  D.phx.[,,k] <- diag(phx.[,k])
  V.h[,,k] <- (N.h..[k]^2)/(N.h..[k]-
1)*(kronecker((D.ph.x[,,k]-ph.x[,k]%*%t(ph.x[,k])),
#covariance matrix

(D.phx.[,,k]-phx.[,k]%*%t(phx.[,k])))

#Pieces that will be summed for calculating test statistic
```

```
  piece1[k,] <- t(n.h[,k]-m.h[,k])%*%t(A.h)
  piece2[,,k] <- (A.h%*%V.h[,,k]%*%t(A.h))
  piece3[,k] <- A.h%*%(n.h[,k]-m.h[,k])
}

#Clearing NANs
piece1[!is.finite(piece1)] <- 0
piece2[!is.finite(piece2)] <- 0
piece3[!is.finite(piece3)] <- 0

#Sums used in Q.GMH test statistic
sum1 <- matrix(colSums(piece1),1,ncol(piece1))
sum2 <- apply(piece2,c(1:2),sum)
sum3 <- rowSums(piece3)

Q.GMH <- sum1%*%(solve(sum2))%*%sum3    #Test statistic
p.value <- pchisq(Q.GMH, df=((groups.R-1)*(cats.C-1)),
lower.tail=F)  #P-value

return(list(Q.GMH=Q.GMH, p.value=p.value))
}
```

```
################################
Liu-Agresti Function
################################

LA <- function(table.data){

  #Extracting info from contingency table
  stratum <- dim(table.data)[3]
  steps <-(dim(table.data)[2])-1
  response.level <- (dim(table.data)[2])

  #Assigning vectors for odds ratio computations
  A.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  B.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  C.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  D.jk <- matrix(c(0), nrow=steps,ncol=stratum)
  N.jk <- matrix(c(0), nrow=steps,ncol=stratum)

  alpha.jnumerator <- matrix(c(0), nrow=steps,ncol=stratum)
  alpha.jdenominator <- matrix(c(0), nrow=steps,ncol=stratum)
  la.matrix <- matrix(c(0),21,4)

##Cumulative category odds ratio (for Liu-Agresti statistic)
  for(k in 1:stratum) {

    for(j in 1:steps){

        #parts of odds ratio being created
        A.jk[j,k] <- sum(table.data[1,(j+1):(steps+1),k])
        B.jk[j,k] <- sum(table.data[1,1:j,k])
        C.jk[j,k] <- sum(table.data[2,(j+1):(steps+1),k])
        D.jk[j,k] <- sum(table.data[2,1:j,k])
```

```
        #total respondents
        N.jk[j,k] <- (A.jk[j,k] + B.jk[j,k] + C.jk[j,k] +
D.jk[j,k])

        #odds ratio numerator and denominator
        alpha.jnumerator[j,k] <-
(A.jk[j,k]*D.jk[j,k])/N.jk[j,k]
        alpha.jdenominator[j,k] <-
(B.jk[j,k]*C.jk[j,k])/N.jk[j,k]

        alpha.jnumerator[!is.finite(alpha.jnumerator)] <- 0
        alpha.jdenominator[!is.finite(alpha.jdenominator)]
<- 0
      }
    }

    LA.numerator <- sum(alpha.jnumerator[,])
    LA.denominator <- sum(alpha.jdenominator[,])
    LA.OR <-  LA.numerator/LA.denominator

    LA.stat <- round(log(LA.OR),3)

###STANDARD ERROR COMPONENTS######

LA.OR <- (1/LA.OR)

#Set variable values
V1 <- 0 #Numerator
V2 <- 0 #Denominator

#Compute components of LA common odds ratio
```

```
for(k in 1:stratum){

  #Set counters to zero
  T1 <- 0 #Reference count
  T2 <- 0 #Focal count

  #Compute stratum-level frequencies
  for(j in 1:response.level){

    T1 = T1 + table.data[1,j,k]
    T2 = T2 + table.data[2,j,k]
  }

  #Compute the variance
  if(T1 >0 && T2>0 && LA.OR>0) {

    CR1 <- 0 #Cumulative reference 1
    CF1 <- 0 #Cumulative focal 1

    for(j1 in 1:(response.level-1)){

      CR1 <- CR1 + table.data[1,j1,k]
      CF1 <- CF1 + table.data[2,j1,k]
      CR2 <- 0 #Cumulative reference 2
      CF2 <- 0 #Cumulative focal 2

      for(j2 in 1:j1){

        CR2 <- CR2 + table.data[1,j2,k]
        CF2 <- CF2 + table.data[2,j2,k]

        #Part 1
```

```
        V3 <- T1 * T2 / ((T1+T2) ^ 2)

        #Part 2
        V4 <- LA.OR * (T1 - CR1) * (CF2) / T1

        #Part 3
        V5 <- 1 + (LA.OR - 1) * CF1/ T2

        #Part 4
        V6 <- CR2 * (T2 - CF1) / T2

        #Part 5
        V7 <- LA.OR - (LA.OR - 1) * (CR1 / T1)

        #Variance component
        V8 <- V3 * ((V4 * V5) + (V6 * V7))

        if(j1 == j2){
          V1 <- V1 + V8
        }
        else{
          V1 <- V1 + 2 * V8}
      }

      #Compute the denominator of the variance
      V2 <- V2 + CR1 * (T2 - CF1) / (T1 + T2)

    }
  }
}

####Compute Standard error
```

```
if(V2 > 0 && LA.OR > 0){
  SE.value <- round(sqrt(V1 /(V2^2)),3)
}
  else{
    SE.value <- NA
  }

Z.value<- round((LA.stat/SE.value),3)
p.value <- round(2 * pnorm(-abs(Z.value)),3)

la.matrix[i,] <-  c(LA.stat,SE.value,Z.value,p.value)

return(list(LA.stat=LA.stat,Z.value=Z.value,SE.value=SE.valu
e,p.value=p.value))
} #closes function




########################################
Benjamini-Hochberg Function
########################################
BH <- function(pvalue.vector, step.value){

  length.pvals <- length(pvalue.vector)
  BH.pvalues <- rep(0, times=length(pvalue.vector))

 if(length(pvalue.vector) > 20){
  for(i in seq(1,(length.pvals), by=step.value)){

      step.vector <- (pvalue.vector[i:(i+step.value-1)])
      j <- step.value:1L
```

```
      o <- order(step.vector,decreasing=TRUE)
      ro <- order(o)
      BH.pvalues[i:(i+step.value-1)] <- pmin(1,
cummin(step.value/j * step.vector[o]))[ro]
  }
 }
    else{
      no.items <- length(pvalue.vector)
      i <- no.items:1L
      o <- order(pvalue.vector,decreasing=TRUE)
      ro <- order(o)
      BH.pvalues <- pmin(1, cummin(no.items/i *
pvalue.vector[o]))[ro]

    }
    return(BH.pvalues)
  }
```

```
####################################
Dunn-Bonferroni Function
####################################

Bonf <- function(pvalue.vector, step.value){
  length.pvals <- length(pvalue.vector)
  bonferroni.pvalues <- rep(0, times=length(pvalue.vector))

  if(length(pvalue.vector) > 20){
  for(i in seq(1,(length.pvals), by=step.value)){

    bonferroni.pvalues[i:(i+step.value-1)] <- pmin(1,
step.value*(pvalue.vector[i:(i+step.value-1)]))
    }
  }
  else{
      no.items <- length(pvalue.vector)
      bonferroni.pvalues <- pmin(1, no.items*pvalue.vector)
    }

 return(bonferroni.pvalues)
}


##############################
##HOLM FUNCTION
##############################

Holm<- function(pvalue.vector, step.value){
```

```
  length.pvals <- length(pvalue.vector)
  Holm.pvalues <- rep(0, times=length.pvals)

  if(length(pvalue.vector) > 20){
    for(i in seq(1,(length.pvals), by=step.value)){

      step.vector <- (pvalue.vector[i:(i+step.value-1)])
      j <- seq_len(length(step.vector))
      o <- order(step.vector)
      ro <- order(o)
      Holm.pvalues[i:(i+step.value-1)] <- pmin(1,
cummax((step.value - j + 1L)* step.vector[o]))[ro]
    }
  }
  else{
    no.items <- length(pvalue.vector)
    i <- seq_len(no.items)
    o <- order(pvalue.vector)
    ro <- order(o)
    Holm.pvalues <- pmin(1, cummax((no.items - i + 1L) *
pvalue.vector[o]))[ro]
  }
  return(Holm.pvalues)
}
```

```
############################################################
CALCULATING TRUE AND FALSE POSITIVES (POWER AND TYPE I ERROR)
FOR ITEMS OF EACH TEST
############################################################
#
TFP.items <- function(items, pvalue.items, ssl.items, cond){

  truep.items <- rep(0, times=(items))
  falsep.items <- rep(0, times=(items))
  reject.items <- rep(0, times=(items))

  #Key for items without DIF
  if(cond == 1 | cond == 4 | cond == 7 | cond == 10 | cond
== 13 | cond == 16 |
       cond == 19 | cond == 22 | cond == 25 | cond == 28 |
cond == 31 | cond == 34 |
       cond == 37 | cond == 40 | cond == 43 | cond == 46){

  dif.items <- rep(0,20)
  nodif.items <- rep(1,20)
  }
```

```
  #Key for items with DIF
  if(cond == 2 | cond == 3 | cond == 5 | cond == 6 | cond ==
8 | cond == 9 | cond == 11 |
       cond == 12 | cond == 14 | cond == 15 | cond == 17 |
cond == 18 | cond == 20 |
       cond == 21 | cond == 23 | cond == 24 | cond == 26 |
cond == 27 | cond == 29 |
       cond == 30 | cond == 32 | cond == 33 | cond == 35 |
cond == 36 | cond == 38 |
       cond == 39 | cond == 41 | cond == 42 | cond == 44 |
cond == 45 | cond == 47 | cond == 48){

  dif.items <- append((c(1,1)), (rep(0,18)))
  nodif.items <- append((c(0,0)), (rep(1,18)))
  }

  for(n in 1: (items)){

    if(length(ssl.items)==1){
      #True positives for items
      ifelse(dif.items[n]==1 && pvalue.items[n] < .05,
truep.items[n] <- 1, truep.items[n] <- 0)
      #False positives for items
      ifelse(nodif.items[n]==1 && pvalue.items[n] < .05,
falsep.items[n] <- 1, falsep.items[n] <- 0)
      #Reject rates for items
      ifelse(pvalue.items[n] < .05, reject.items[n] <- 1,
reject.items[n] <- 0)
    }

    else{
```

```
      #True positives for items
      ifelse(dif.items[n]==1 && ssl.items[n]==1,
truep.items[n] <- 1, truep.items[n] <- 0)
      #False positives for items
      ifelse(nodif.items[n]==1 && ssl.items[n]==1,
falsep.items[n] <- 1, falsep.items[n] <- 0)
      #Reject rates for items
      ifelse(ssl.items[n]==1, reject.items[n] <- 1,
reject.items[n] <- 0)
    }
  }

  #for items per test
  power.items <- sum(truep.items)/sum(dif.items)
  t1err.items <- sum(falsep.items)/sum(nodif.items)

return(list(truep.items=truep.items,falsep.items=falsep.item
s,reject.items=reject.items,
              power.items=power.items,
t1err.items=t1err.items))
}
##############################################################
#
CALCULATING TRUE AND FALSE POSITIVES (POWER AND TYPE I ERROR)
FOR ITEM STEPS OF EACH TEST
##############################################################
#

TFP.steps <- function(item.steps, items, pvalue.steps,
cond){

  truep.steps <- rep(0, times=(item.steps*items))
```

```
  falsep.steps <- rep(0, times=(item.steps*items))
  reject.steps <- rep(0, times=(item.steps*items))
  power.steps <- rep(0, times=(items))
  t1err.steps <- rep(0, times=(items))

 # Key for steps with or without DIF/DSF
  if(cond== 1 | cond == 7 | cond == 13 | cond == 19 | cond
== 25 | cond == 31 |
       cond == 37 | cond == 43){

    dsf.steps <- rep(0,40)
    nodsf.steps <- rep(1,40)
  } #closes if loop cond 1...

  if(cond == 2 | cond ==3  | cond == 8 | cond == 9 | cond ==
14 | cond == 15 |
       cond == 20 | cond == 21 | cond == 26 | cond == 27 |
cond == 32 | cond == 33 |
       cond == 38 | cond == 39 | cond == 44 | cond == 45){

    dsf.steps <- append((c(1,1,1,1)), (rep(0,36)))
    nodsf.steps <- append((c(0,0,0,0)), (rep(1,36)))

  } #closes if-loop cond 2,3...

  if(cond == 4 | cond == 10 | cond == 16 | cond == 22 | cond
== 28 | cond == 34 |
       cond == 40 | cond == 46){

    dsf.steps <- rep(0,60)
    nodsf.steps <- rep(1,60)
  } #closes if-loop cond 4
```

```
  if(cond == 5 |cond == 6 | cond == 11 | cond == 12 | cond
== 17 | cond == 18 |
       cond == 23 | cond == 24 | cond == 29| cond == 30 |
cond == 35 | cond == 36 |
       cond == 41 | cond == 42 | cond == 47 | cond == 48){

    dsf.steps <- append((c(1,1,0,1,1,0)), (rep(0,54)))
    nodsf.steps <- append((c(0,0,1,0,0,1)),(rep(1,54)))
  } #closes if-loop cond 5,6

 for(m in 1:(item.steps*items)){
    #True positives for steps
    ifelse(dsf.steps[m]==1 && pvalue.steps[m] < .05,
truep.steps[m] <- 1, truep.steps[m] <- 0)
    #False positives for steps
    ifelse(nodsf.steps[m]==1 && pvalue.steps[m] < .05,
falsep.steps[m] <- 1, falsep.steps[m] <- 0)
    #Reject rates for steps
    ifelse(pvalue.steps[m] < .05, reject.steps[m] <- 1,
reject.steps[m] <- 0)

} #closes pvalues m forloop
```